FedRCIL: Federated Knowledge Distillation for Representation based Contrastive Incremental Learning

Athanasios Psaltis*[1,2], Christos Chatzikonstantinou*[1], Charalampos Z. Patrikakis[2], and Petros Daras[1]

[1]Centre for Research and Technology Hellas, Thessaloniki, Greece
[2]Dept. of Electrical and Electronics Engineering, University of West Attica, Athens, Greece
{chatziko, daras}@iti.gr {apsaltis, bpatr}@uniwa.gr

Abstract

The present work proposes a holistic approach to address catastrophic forgetting in the field of computer vision during the process of incremental learning. More specifically, it suggests a series of steps for effective learning of models in distributed environments, based on extracting meaningful representations, modeling them into actual knowledge, and transferring it through a continual distillation mechanism. Additionally, it introduces a federated learning algorithm tailored to the problem, eliminating the need for central model transfer, by proposing an approach based on multi-scale representation learning, coupled with a Knowledge Distillation technique. Finally, inspired by the current trend, it modifies a contrastive learning technique combining existing knowledge with previous states, aiming to preserve previously learned knowledge while incorporating new knowledge. Thorough experimentation has been conducted to provide a comprehensive analysis of the issue at hand, highlighting the great potential of the proposed method, achieving great results in a federated environment with reduced communication cost and a robust performance within highly distributed incremental scenarios.

1. Introduction

The ever increasing number of smart devices leads to a rapid expansion in the amount of data being exchanged, presenting a substantial challenge, especially when handling distributed data. This becomes increasingly challenging when new labeled or unlabeled data are continuously introduced to the system. As a result, the demand for systems that can effectively exploit these data to adapt to new conditions, *i.e*. tasks, while minimizing costs becomes even more pronounced. Several researchers dedicated resources to find reliable solutions [8], but despite the vast amount of published research on the topic, the challenge of learning new knowledge without forgetting, particularly in the context of distributed datasets, still poses significant challenges.

In the realm of computer vision, to achieve truly robust recognition performance and maintain the capacity to recognize previously seen entities (*e.g*. objects or scenes) when encountering new classes or datasets, it is crucial to tackle several pivotal challenges. These include managing data heterogeneity and model bias effect, communication and synchronization, privacy and security concerns, as well as ensuring scalability and efficiency in processing and training distributed datasets to accommodate the increased complexity [24]. To tackle these, research efforts primarily concentrated on utilizing combinations of regularization techniques and data synthesis methods to mitigate or alleviate catastrophic forgetting, while network expansion techniques and rehearsal-based approaches were explored to expand network capacity and prioritize samples based on their importance for learning, respectively [29]. Recent advancements in lifelong learning approaches [4, 37, 5] have gained attention in incremental learning computer vision research, reshaping the way visual analysis is conducted. These techniques enable to learn generic learning patterns that can adapt to new tasks and data, allowing the model to continually improve its learning efficiency and adaptability. By leveraging knowledge from previous tasks to facilitate learning the new ones, these methods enable more robust and coherent incremental learning.

While there has been extensive research on applying incremental learning to centralized data settings, the exploration of its application to federated datasets and tasks, is still relatively limited, due to data privacy concerns, system heterogeneity constraints, dynamic data distribution, *etc*. Under a realistic scenario, end-users, or the nodes of the

*Equal Contribution

system, constantly generate new data that neither necessarily belong to previously known categories nor follow a specific distribution. In an attempt to understand the problem and train a general model, older versions of models and data are maintanined, which constantly push the limits of system requirements, inevitably leading to a sudden significant decrease in their performance in previous tasks. Research groups have devoted resources for approaches specifically designed for federated settings, leveraging the power of incremental learning while respecting the decentralized nature of data. These approaches often involve a range of learning techniques, such as knowledge distillation, lifelong learning, and hybrid approaches to achieve better performance in preserving previous knowledge and learning new knowledge incrementally [10, 9, 11, 23].

Through the application of a combination of techniques such as representation learning, knowledge distillation, and contrastive learning (CLR) in our approach, it becomes feasible to address the challenges posed by dynamic, heterogeneous, and fragmented data. In a more specific context, the generation of representations that encompass knowledge from diverse levels of the model results in a robust feature extractor with enhanced capabilities. When combined with the Knowledge Distillation approach, this integration facilitates effective generalization at a federated level, allowing for broader applicability and improved performance. For incarnating the task incremental step, we leveraged CLR [6] to distinguish optimal representations by contrasting both global and local representation samples. The proposed methodology aimed to enhance the overall quality of representations by creating a latent space where global representations captured shared knowledge across different tasks, while task-specific representations were well-separated, enabling effective discrimination and understanding of individual tasks.

The main contributions of this work are summarized as follows:

(a) **A novel Federated Incremental Learning scheme is introduced** that utilizes multi-level representation learning coupled with rehearsal-based knowledge distillation approaches that support FL optimization algorithms and CRL techniques in an end-to-end learning manner. The holistic nature of our approach ensures a robust and scalable solution for incremental learning in federated environments.

(b) **A rehearsal-based knowledge distillation technique to transfer this enriched knowledge to neighboring nodes.** Through knowledge distillation, the learned insights and patterns from the well-mixed representations are effectively transferred and shared across the federated network.

(c) **A contrastive-based learning algorithm that utilizes mixed representations to retain knowledge gained from previous tasks of the incremental scenario.** By combining features from different tasks, our algorithm encourages the model to learn and preserve valuable information across different task domains.

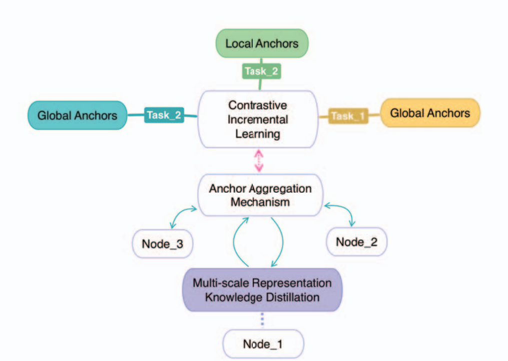
Local Anchors
Task_2
Global Anchors
Task_2
Contrastive Incremental Learning
Task_1
Global Anchors
Anchor Aggregation Mechanism
Node_3
Node_2
Multi-scale Representation Knowledge Distillation
Node_1

Figure 1: The proposed Federated Learning (FL) scheme utilizes a centralized server to create and share the aggregated global representations, while several local nodes participate asynchronously in the training process.

(d) **Validate the effectiveness of the proposed approach through extensive comparisons on one of the broadest publicly available benchmarks,** demonstrating its superiority and robustness across a broad spectrum of FL settings.

The remainder of this paper is organized as follows: Related work is reviewed in Section 2. The proposed Federated Incremental learning scheme is presented in Section 3. Implementation details along with experimental results are discussed in Section 4 and conclusions are drawn in Section 5.

2. Related Work

2.1. Knowledge distillation

Knowledge distillation is a technique that aims to efficiently transfer information from a large model (known as teacher model) to a smaller one (known as student model). The teacher model guides the student model to achieve a better performance through an iterative learning process. Knowledge distillation is widely used for model deployment on resource constrained devices. The concept of distillation learning is introduced in [16], using soft outputs of the teacher model in order to guide the training of the smaller model. Moreover, a distillation loss is introduced, combined with the cross entropy loss to strike a balance between data fitting and mimicking the teacher. In most recent works, Zhao *et al.* [39] divides the knowledge distillation in two parts, namely the target and the non-target. The target knowledge distillation part is a binary logit distillation for the target class and the non-target knowledge distillation part is a multi-category logit distillation for non-target classes. In [18], the deviation between the predictions of the teacher and the student model is addressed. A correlation-based loss is introduced to capture inter-class and intra-class relations from the teacher. The technique of knowledge distillation has also extended to the field of FL as described in 2.3

2.2. Incremental learning

The Incremental Learning problem has been the focus of various studies, encompassing different granular settings. Typically, it can be broadly categorized into three types, depending on the specific characteristics of the incremental scenario, namely task-incremental learning, class-incremental learning, and domain-incremental learning [2]. The first two share a similar setting in which new classes are introduced in new tasks. However, the key distinction between them lies in the inference stage. Unlike other incremental learning scenarios, where new tasks introduce new classes, domain-incremental learning addresses the challenge of adapting to shifts in data distribution or domains while preserving the existing label space. Each scenario presents unique challenges, such as avoiding catastrophic forgetting, handling domain shifts, or managing imbalanced data distributions, and requires tailored approaches to ensure effective continual learning [28]. Therefore, research on incremental learning covers a wide range of approaches, including regularization-based methods, rehearsal and replay techniques, knowledge distillation, network expansion, and hybrid approaches that combine multiple strategies [3, 34, 26, 1, 40, 36, 27].

CRL serves as a potent instrument within the realm of incremental learning. Its fundamental aim revolves around crafting representations that induce a tendency for similar instances to congregate in the embedding space (*i.e*. positive pairs), while simultaneously propelling dissimilar instances to be dispersed at a greater distance (*i.e*. negative pairs). Inspired by the typical CRL framework in [6], recent breakthroughs in incremental learning argue that contrastively learned representations are robust against the catastrophic forgetting [32, 4, 13, 37, 5], and could be transferred better to unseen tasks.

In recent studies, several approaches have been proposed to address the problem of catastrophic forgetting in different domains. Cha *et al*. in [4] propose a rehearsal-based continual learning algorithm that utilizes CRL and self-supervised distillation to learn and maintain transferable representations, leading to improved performance in image classification tasks. In a similar attempt, authors in [13], utilize instance-level and class-level contrastive losses, along with knowledge distillation and a spatial group-wise enhanced attention mechanism, to maintain the inner-class assignment information and alleviate catastrophic forgetting. A novel incremental learning framework is proposed in [32], which utilizes contrastive one-class classifiers to address catastrophic forgetting in class incremental learning. Additionally, [5] extends contrastive self-supervised learning to be primarily based on exemplars and applicable to both labeled and unlabeled data, enabling few-shot class incremental learning.

This theory paves the way for an incremental learning task, wherein the network endeavors to diminish the distance between the current and the previous task instances in the latent space, while concurrently increasing the space between all the other instances within the dataset, effectively achieving a balance between stability and plasticity. In line with this concept, the network aims to grasp the task-specific representation of the instances in a way that improves their distinguishability in the embedding space. This facilitates the identification and differentiation of individual instances, making the overall process simpler.

2.3. Federated learning

In contrast to traditional centralized machine learning (ML) techniques, FL employs a training approach where an algorithm is trained through multiple independent sessions, with each session using its own distinct dataset. FL enables multiple actors to collaboratively train a shared and resilient ML model without the need to centralize their respective data. Through this approach, FL effectively addresses concerns related to data privacy, security, and authorization while allowing for the utilization of diverse and heterogeneous data sources.

Initially, the research on FL has primarily concentrated on enhancing communication efficiency and expediting model updates. The groundbreaking work by McMahan *et al*. [30] introduces a novel concept of averaging local stochastic gradient descent updates (known as FedAvg) to increase the overall amount of information used of each client during communication rounds. To overcome challenges like low device participation and non-independent and identically distributed (Non-IID) local data, several studies have explored the use of online knowledge distillation approaches. A novel approach for federated multi-task distillation is introduced in [35], while Wu *et al*. [33] presented a communication-efficient FL method, utilizing adaptive mutual knowledge distillation and dynamic gradient compression to reduce communication costs. Similarly, Li *et al*. [25] introduces a unified algorithmic framework for Federated Distillation (FD), employing active data sampling to reduce communication overhead.

Towards this direction, a recently introduced technique has emerged, named FLD (Federated Learning Distillation), which takes a different approach, by exchanging and aggregating client model outputs. This alternative methodology offers distinct advantages such as reduced communication costs, flexibility in model architecture, and improved handling of Non-IID data distribution. Jeong *et al*. [19] propose an FLD scheme, where clients upload per label averaged soft targets to train a conditional Generative Adversarial Network (GAN). Li *et al*. [22] introduce a common dataset accessible to all clients, training them on the public dataset before their private data. Gong *et al*. [12] introduce one-shot distillation where client models are fully trained and then distilled to the server using attention maps per class. These approaches combine distillation and aggregation mechanisms to facilitate FL, while also considering the integration of public or shared datasets.

Dealing with catastrophic forgetting in FL poses unique challenges that have not been extensively explored compared to other learning settings. To address this gap, researchers have proposed a set of models [10, 9, 11] specifically designed for federated scenarios. These models incorporate techniques such as class-aware gradient compensation, semantic distillation, adaptive class-balanced pseudo labeling, and forgetting-balanced semantic compensation. Additionally, a recent study introduced a local model contrastive loss [23] to enhance individual party training. These approaches effectively mitigate forgetting and provide solutions for catastrophic forgetting in the FL context.

While there has been notable progress in Federated Incremental Learning, there is still untapped potential in exploring strategies that leverage the best solutions from the literature on knowledge distillation and incremental learning and adapt them to the federated setting. The existing research presents opportunities to develop novel techniques that effectively address challenges in that field.

3. Proposed method

3.1. Problem statement

In a federated system, the data are inherently localized to individual clients, and sharing them with other clients is strictly prohibited. The objective of this study is to learn from a continuously evolving stream of data that introduces new classes in highly distributed environments. The data are divided into a sequence of non-overlapping training tasks, each representing a step of incremental learning. The final goal is to continually develop a classification model that incorporates knowledge from both current and previous tasks across various federated nodes. After each task, the model's performance is evaluated on all the classes encountered so far, which is the union of classes from all previous tasks.

The proposed methodology involves a federated incremental learning scheme that combines multi-level representation learning, knowledge distillation approaches coupled with FL optimization algorithms, and CRL techniques to address the challenges of evolving data domains, as depicted in Figure 1. For the cultivation of individual local clients, the CIFAR-100[21] dataset is deployed, partitioned in both a balanced and unbalanced manner to facilitate experiments that cater to both IID and Non-IID conditions.

This work follows the common practice of using exemplar sets for incremental learning in computer vision. Exemplars are representative instances of known classes, selected from the training set. Instead of random sampling, the herding strategy is employed to choose the most representative exemplars for each class. Mean anchor images from the previous rounds are computed, and the exemplar set is formed by selecting class images that approximate these mean anchors [31].

3.2. Federated Knowledge Distillation

3.2.1 Local Supervision Representation Learning

The local clients undergo supervised training using the CIFAR-100 dataset, with each client being assigned a specific segment of the dataset. This assignment remains consistent throughout all federated rounds, and the clients receive training on their allocated segments. The training subset of the dataset is utilized for model training, while the validation subset is used to evaluate and authenticate the performance of each client's trained models. Apart from each client's training and validation set, a communal dataset exists that is accessible to all clients. To train the clients' networks, a classification head was added to map the feature vectors to the total number of classes in the dataset; 100 in the case of CIFAR-100. The Cross Entropy was deployed as the loss function for the training process, and its formula is shown below:

$$L(\theta) = -\tfrac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} y_{mk} \log p_{mk} \qquad (1)$$

where M is the total number of images, K the total number of classes, y_{mk} the label of the m, k image and p_{mk} the probability of the class. The stochastic gradient decent (SGD) [20] algorithm was chosen as the optimization strategy.

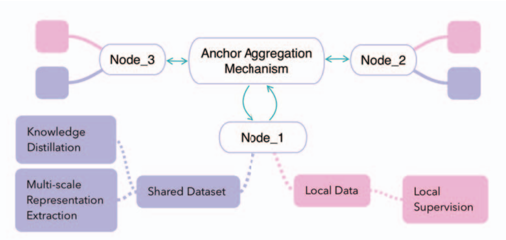

Figure 2: The proposed Federated Multi-scale Representation Knowledge Distillation scheme.

One drawback of the federated distillation scheme is the small amount of information transmitted from the clients to the data, since only the model's output is transmitted to the central server (*i.e*. last fully-connected layer). In order to mitigate this drawback, a scheme of multi-scale knowledge distillation is adopted inspired by the work of [14]. More specifically, apart from the model's output each client transmits some intermediate outputs (*i.e.* hidden representation), where the number and the position of the intermediate outputs can vary depending on the problem.

Different-level features can be extracted, let $z_l^N = f_l^N(\cdot), l \in (1, ..., L)$, where l are the different model layers, z_l^N is the intermediate output at layer l for the client N, with size $H * W * C$ and $f_l^N(\cdot)$ is the subnetwork for feature extraction up to the layer l. Then, the final form of the feature map can be obtained by concatenating the $W*C$ height-pooled slices and the $H * C$ width-pooled slices for h_l^N:

$$\Phi(z_{l_{j-1}}^N) = \left[\tfrac{1}{W} \textstyle\sum_{w=1}^{W} z_{l_{j-1}}^N [:, w, :] \middle|\middle| \tfrac{1}{W} \textstyle\sum_{h=1}^{H} z_{l_{j-1}}^N [h, :, :] \right] \quad (2)$$

where $[\cdot|\,|\cdot]$ denotes concatenation over the channel axis, N is the number of clients, j is the federated round and $\Phi(\cdot)$ is the network function. The intermediate outputs are calculated over the communal dataset, therefore all clients extract features from the same subset. After all intermediate outputs of all clients are extracted, they are averaged per client, $\Phi(z_l) = \frac{1}{N} \sum_{n=1}^{N} \Phi(z_{l_{r-1}}^N)$. At the beginning of the next federated round, for each client, the intermediate outputs are employed as soft labels in order to minimize the Euclidean distance between the mean averaged representation of the clients and each client's local representation. The multi-scale knowledge distillation term is calculated as:

$$\mathcal{L}_{multi-scale_{KL}} = \frac{1}{L} \sum_{l=1}^{L} \Big| |\Phi(z_{l_j}^N) - \Phi(z_{l_{j-1}}) \Big| | \quad (3)$$

3.2.2 Global Distilled Supervision

As previously mentioned, a communal dataset is available and accessible to all clients, following the common practice in the literature ([22], [17], [7]). This communal dataset serves as a reference point to address data heterogeneity among the clients in the FL scheme. At the end of each federated round, every client generates model outputs for each sample in the common dataset. These outputs are then transmitted to the server side and averaged across all clients (illustrated in Figure 2), similar to the approach described in 3.2.1. In the subsequent federated round, each client performs knowledge distillation using the global anchors derived from the common dataset, thus, ensuring knowledge alignment between the clients. The knowledge distillation term is calculated as follows:

$$\mathcal{L}_{KL} = \frac{1}{L} \sum_{l=1}^{L} \big| |\Phi(o_j^N) - \Phi(o_{j-1}) \big| | \quad (4)$$

where o_r^c is the model output of client N at the federated round j.

3.3. Incremental Learning

The proposed federate incremental scheme draws inspiration from MOON [23], which is a simple and effective approach based on FedAvg with lightweight modifications in the local training phase. However, there are significant differences in our approach. We focus solely on representation learning at each federated round using knowledge distillation. Our approach involves extracting stronger representations through a robust aggregation algorithm. In contrast to MOON, we incorporate CRL, which is typically used for visual representations, on the aggregated global anchors and their respective local versions (as presented in Figure 1). This allows us to decrease the distance between the representation learned by the local model and the global aggregated representation of the previous task(s), while increasing the distance between the representation learned by the local model and the global aggregated representation of the previous federated round of the same task. By incorporating these elements, our approach enhances the learning procedure and leverages the power of CRL in the federated setting. Moreover, the global aggregated representation of all previous tasks is employed to further enhance the learning procedure.

In the federated incremental architecture, the primary server undergoes training via a contrastive approach on the dataset $X_m = (x_m^i, y_m^i), i = 1, ...N$, wherein the labels are disregarded. Let us assume that node N_i is performing the local training. Each image from the said dataset serves as the anchor image x_m^i. All of these images are subsequently passed through the network function $\Phi(\cdot)$, and are then projected into the latent space for the implementation of CLR as delineated in [6]. During the initialization phase, each node receives the global aggregated anchors a_t from the server, which are common for all clients at the current round. During this process, for each input image x_m^i, we extract the representation of x_m^i, following the federated knowledge distillation approach defined above, from the current global aggregated anchors from the previous federated round a_t^{j-1} as $z_t^{j-1} = \Phi a_t^{j-1}(x_m^i)$, the representation of x_m^i from the aggregated global anchors of the previous incremental round a_{t-1} as $z_{t-1} = \Phi a_{t-1}(x_m^i)$, and the representation of x_m^i from the current local anchor being updated a_t^j as $z_t^j = \Phi a_t(x_m^i)$. The feature maps are harnessed for the calculation of the contrastive loss. The formula of the loss is indicated below:

$$a_{con} = -\log \frac{\exp\left(sim(z_t^j, z_{t-1})/T\right)}{\exp\left(sim(z_t^j, z_{t-1})/T\right) + \exp\left(sim(z_t^j, z_t^{j-1})/T\right)} \quad (5)$$

where $\mathcal{T}$ is the temperature that is a scaling factor used to control the concentration of the output distribution, affecting the hardness of the positives in the CLR framework. Since the global model is expected to generate better representations, our objective is to minimize the distance between z_t^j and z_t^{j-1}, indicating that the local model aligns with the previous incremental round's global anchor's representations. Additionally, we aim to maximize the distance between z_t^j and z_{t-1}, indicating that the updated local anchors are diverging from the previous round's global anchor's representations and retaining their own learned representations. Through the implementation of contrastive loss, the network strives to learn representations that induce a proximity between the local image and the positive anchor global image in the embedding space, whilst ensuring a separation between the local image and the previous round's global anchor image. The representations learned in this manner would have the capacity to segment the latent space in accordance with the context, without the true comprehension of the task.

3.4. FedRCIL Algorithm

In this section, the complete flow of the described procedure is presented. The entire process is divided into three main algorithms, the Algorithm 1 describes the required steps for training and updating the network in the main server, while the other two define the actions for training and updating the network in each local client (*i.e.* Algorithm 2 for local and Algorithm 3 for global supervision respectively).

4. Experimental Results

4.1. Dataset settings

In this section, experimental results from the application of the proposed FedRCIL scheme, using the CIFAR-100 image classification dataset as a benchmark, are presented. The CIFAR-100 dataset was adapted to represent both IID and Non-IID scenarios, capturing different FL settings. In the IID setup, the data was balanced, while the Non-IID setup represented an extreme case of data imbalance. Like previous studies [38], Dirichlet distribution is utilized to generate the Non-IID data partition among parties, using a concentration parameter β. Both datasets consisted of 100 categories, and a total of 10 participating nodes were involved, with images distributed among them.

4.2. Implementation Details

4.2.1 Architecture and Parameters

The architecture of the proposed method employs a distributed framework, where each local client utilizes ResNet56 [15]. The ResNet56 model is chosen for its fast training and satisfactory results. At each local client, the final fully connected layer of the network is replaced with a projection head to facilitate the Incremental CRL task, allowing the transition of ResNet feature maps to a new embedding space that supports the CLR scheme. Meanwhile, the local clients undergo supervised training using the proposed multi-loss scheme, involving three intermediate outputs corresponding to the outputs of three ResNet layers. The SGD method is employed with a learning rate of 0.1, the batch size of the system is 64, the m_u and $\mathcal{T}$ are set equal to 0.5 and m_c equals to 0.1 . The local models are trained for 300 epochs to ensure comprehensive learning and convergence. Python 3.7 and PyTorch (version 1.7.0) environments are employed for the impelemntation of the deep learning models. The code is available at `https://github.com/chatzikon/FedRCIL`.

Algorithm 1 FedRCIL Algorithm

```
Require: $T$ is the number of communication rounds, $\mathcal{N}$ is the total number of clients, $\theta_l^i$ represents the parameters of the local models, $\mathcal{D}_l^i$ are the separate datasets for the local clients, $\mathcal{D}_{ex}$ is the dataset of the exemplar set, $\mathcal{D}_{test}$ is the common testing dataset, and $\eta$ is the learning rate.
for each $i$ from 1 to $N$ do
    Initialize local models $\theta_l^i$
    Prepare local datasets $D_l^i$
end for
Prepare exemplar set $D_{ex}$
Prepare common global dataset for evaluation $D_{test}$
Server executes:
for each Incremental round $t = 0, 1, 2, \dots$ do
    for each Federated round $j = 0, 1, 2, \dots$ do
        for each client in parallel do
            $\theta_{l_j} \leftarrow$ ClientUpdate($D_l, \theta_{l_j}$)
            $z_{t_j}, .. \leftarrow$ Extract representations from ($\theta_{l_j}$)
            Send representation to server
        end for
        $a_{t_j} \leftarrow$ Aggregate Anchor representations
        Distribute the updated global Anchors
        for each client in parallel do
            $\theta_{l_{j+1}} \leftarrow$ KDUpdate($D_{ex}, \theta_{l_j}, a, m_u = 0.5$)
        end for
    end for
    Validate the updated distillated models on $D_{test}$
end for
```

Algorithm 2 ClientUpdate Function

```
ClientUpdate($D_l$, $\theta$):                ▷ Run on specific client
for each local epoch $i$ from 1 to $E$ do
    for each batch in $D_l$ do
        $\theta_l \leftarrow \theta_l - \eta \nabla L(\theta_l, \text{labels})$   ▷ Update the client model with Eq. 1
    end for
end for
return $\theta_l$
```

Algorithm 3 DistillationUpdate Function

```
KDUpdate($D_{ex}$, $\theta_l$, $a$, $m_u$):    ▷ Run on specific client
for each batch in $D_{ex}$ do
    $l_{mul}$    ▷ Compute Multi Scale loss with Eq. 3
    $l_{dis}$    ▷ Compute Distillation loss with Eq. 4
    $l_{con}$    ▷ Compute constrastive loss with Eq. 5
    $L = m_c * l_{mul} + l_{dis} + m_u * l_{con}$
    $\theta_l \leftarrow \theta_l - \eta \nabla L(\theta_l)$ ▷ Update the client model with aggregated loss $L$
end for
return $\theta_l$
```

4.2.2 Federated setting

The training paradigm for the suggested FL system which comprises a centralized server and 10 local clients, involves 6 federated rounds without incremental learning. In each round, the local clients individually undergo training for 50 epochs without any inter-client communication. The final evaluation of the local models is performed using the test set of the CIFAR-100 dataset. The training set is divided into a common set accessible to all clients (20% of the train set), while the remaining 80% is divided among the clients. In the case of incremental learning, the epochs without any inter-client communication are 10 and the federated rounds are 30. The incremental learning process consists of 5 different tasks, each one with 20 unique classes. Each task has a training period of 6 federated rounds that constitute an incremental round. The common set mentioned above is employed as an exemplar set, with a steady size but with a varying number of samples per class (*i.e.* as new tasks arrive, less samples per class exist), employing the approach mentioned in Section 3.1.

4.2.3 Baselines

To validate the effectiveness of FedRCIL comprehensively, a comparative performance analysis was conducted on two distinct configurations. Initially, benchmark experiments were undertaken within a fully-supervised FedAvg framework, in both IID and Non-IID scenarios. Furthermore, the proposed approach was compared with a scheme of local isolated clients (without communication among them) and with a baseline federated distillation approach.

4.3. Performance Evaluation

This section provides a summary of the results obtained through the application of the proposed scheme in several distinct scenarios under various learning settings. In an attempt to showcase the distinctive characteristics of our architecture, we conducted direct comparisons with the proposed baseline methods, depicted in Table 2. However, it was not possible to compare with other methods from the literature, except for FedAvg, due to the difficulty of adapting state-of-the-art techniques to the specific problem we are addressing.

4.3.1 Representation learning setting

The accuracy results presented in Table 1 investigate various concepts associated with extracting multi-level and multi-scale representations from the local clients. It is shown, that the FLD_{m_B}, where each extra loss is applied to the part of the model before it, back-propagating with layer 1 loss first and layer 3 loss last, achieved the highest accuracy of 38.06%. Similarly, FLD_{m_C}, with losses back-propagating with layer 3 loss first and layer 1 loss last, obtained a slightly lower accuracy of 37.82%. On the other hand, FLD_{m_A}, where all losses apply to the whole network, resulted in the lowest accuracy of 32.03%. These findings highlight the significance of selectively applying additional losses to specific parts of the model and the importance of the direction of back-propagation. The results clearly demonstrate that the proposed method outperforms the conventional cross-entropy loss approaches.

Multi-loss concept	Accuracy
FLD	28.41
FLD_{m_A}	32.03
FLD_{m_B}	38.06
FLD_{m_C}	37.82

Table 1: FLD_{m_A}: All losses apply to the whole network FLD_{m_B} : Each extra loss apply to the part of the model before it, backprop from layer 1 to layer 3 FLD_{m_C} : Each extra loss apply to the part of the model before it, backprop from layer 3 to layer 1

4.3.2 Federated distillation learning setting

In relation to a comparison with the baseline methods, the proposed approach, achieved the highest accuracy of 38.06%, suggesting that leveraging a shared dataset and optimizing multiple objectives simultaneously enhances the learning process. On the other hand, isolated clients where each client trains independently, had the lowest accuracy of 23.29%, of course this was something the we expected, but it highlights the added value of the proposed approach. FL without a common dataset also performed relatively poorly, with an accuracy of 20.61%. Concerning the investigated FLD schemes, the findings suggest choice of layer output, and the presence of a common dataset can influence the effectiveness of distillation. As discernible from Table 2, when juxtaposed with the FedAvg, the proposed approach provides superior results with relative improvement over the baseline of 11.39%. These findings highlight the importance of collaboration and the potential benefits of utilizing shared data and optimized loss functions in FL, ultimately leading to improved accuracy in image classification tasks.

Method	Common set	Accuracy
$FL(FedAvg)$		34.17
Local Isolated Clients		23.29
FLD_l		20.61
FLD_{l-1}	✓	19.71
FLD_l	✓	28.41
$FedRCIL$ (Proposed)	✓	38.06

Table 2: Comparative evaluation with baseline methods and various distillation schemes.

4.3.3 Incremental learning setting

Table 3 presents accuracy results for the proposed method at different m_u settings, which represent the weight of the contrastive loss in the learning process. Utilization of the contrastive learning mechanism, results in significantly higher accuracy, contrary to the case that m_u equals to 0, resulting in a significant drop in accuracy. In particular, the results indicate that the accuracy decreased by around 20% when moving from 1 task to 5 tasks. This significant drop highlights the superior added value of the proposed incremental mechanism. On the other hand, in the other three cases where m_u values are higher (*i.e.* $0.25, 0.5$, and 1), the results remain favorably comparable.

m_u	task=1	task=5
$FedRCIL^{m_u=0}$	29.54	10.96
$FedRCIL^{m_u=0.25}$		26.07
$FedRCIL^{m_u=0.5}$		26.79
$FedRCIL^{m_u=1}$		25.42

Table 3: Experiments with different values for the contrastive loss coefficient m_u.

In an attempt to investigate the impact of knowledge retention on performance, we intentionally increased the number of buffers per task progressively from 1 to 4, as depicted in Table 4. Keeping only the most recent task instance (buffer size $b = 1$) yields the lowest accuracy of 15.76%, indicating its inferior performance. However, as the buffer size increases, accuracy increases significantly, with buffer size $b = 4$ resulting in 26.79%. Buffer size values of 2 and 3 strike a balance between retaining knowledge and model performance, achieving accuracies of 19.52% and 21.12%, respectively. These findings emphasize the importance of managing buffer size in incremental learning; storing an adequate number of previous task instances can positively impact performance, while storing only the most recent one results in the worst accuracy. It is evident that balancing the buffer size is crucial for enhancing knowledge transferability and optimizing the proposed method's effectiveness in incremental learning settings.

Buffer	Accuracy
$FedRCIL^{b=1}$	15.76
$FedRCIL^{b=2}$	19.52
$FedRCIL^{b=3}$	21.12
$FedRCIL^{b=4}$	26.79

Table 4: Experiments with different buffer size (number of previous task models employed as positives at the contrastive learning).

Non-IID data typically contain diverse patterns and variation across different nodes, which inherently lead to challenges in learning a generalisable model. In that context, Table 5 presents accuracy results for three different methods: FLD_c, FLD_{mc}, and the proposed $FedRCIL$, at various beta values ($0.25, 0.5, 0.75$, and 1), which control the non-iidness of the dataset. As beta increases, all methods generally show improved accuracy. The basic FLD_c method achieves the lowest accuracy, ranging from 16.45% ($beta = 0.25$) to 22.10% ($beta = 1$). Introducing multi-loss (FLD_{mc}) leads to higher accuracy across all beta values, ranging from 18.48% ($beta = 0.25$) to 28.79% ($beta = 1$). Notably, the proposed incremental $FedRCIL$ scheme demonstrates promising performance, achieving comparable results of 16.62% with an acceptable decrease of around 10% compared to the IID case. This outcome is particularly impressive considering the challenging evaluation setting involved in incremental learning tasks.

$beta$	0.25	0.5	0.75	1	IID
FLD_c	16.45	20.01	20.64	22.10	28.41
FLD_{mc}	18.48	24.07	25.86	28.79	38.06
$FedRCIL$	10.22	13.29	14.82	16.62	26.79

Table 5: Experiments with Non-IID data, for different $beta$ values. Methods indicated with subscript 'm', and 'c' utilize multi-loss, and common dataset, respectively.

5. Conclusions

This work presented a holistic approach to mitigate catastrophic forgetting and maximizing knowledge retention in computer vision during incremental learning, by integrating CRL, FL and rehearsal-based knowledge distillation techniques. It provides a comprehensive solution for continual learning in scenarios where FL has not been extensively studied, facilitating the learning and preservation of transferable representations. The comprehensive experimentation conducted, provides valuable insights into the proposed learning scheme, particularly in highly distributed scenarios. This research showcases its ability to create robust local representations, effectively replacing the computationally expensive transmission of bulky models, while it also managed to reduce the number of communication rounds by incorporating aggregated knowledge from neighboring nodes and previous tasks, resulting in satisfactory performance levels.

Acknowledgment

This research was supported by the European Commission under contract No. 101073951 LAGO.

References

[1] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 583–592, 2019.

[2] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135:38–54, 2021.

[3] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9516–9525, October 2021.

[5] Daniel T Chang. Exemplar-based contrastive self-supervised learning with few-shot class incremental learning. *arXiv preprint arXiv:2202.02601*, 2022.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Sijie Cheng, Jingwen Wu, Yanghua Xiao, and Yang Liu. Fedgems: Federated learning of larger server models via selective knowledge fusion. *arXiv preprint arXiv:2110.11027*, 2021.

[8] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[9] Jiahua Dong, Yang Cong, Gan Sun, Yulun Zhang, Bernt Schiele, and Dengxin Dai. No one left behind: Real-world federated class-incremental learning. *arXiv preprint arXiv:2302.00903*, 2023.

[10] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10164–10173, 2022.

[11] Jiahua Dong, Duzhen Zhang, Yang Cong, Wei Cong, Henghui Ding, and Dengxin Dai. Federated incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3934–3943, June 2023.

[12] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Ensemble attention distillation for privacy-preserving federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15076–15086, 2021.

[13] Jia-yi Han and Jian-wei Liu. Instance-level and class-level contrastive incremental learning for image classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[14] Ya-nan Han and Jian-wei Liu. Online continual learning via the meta-learning update with multi-scale knowledge distillation and data augmentation. *Engineering Applications of Artificial Intelligence*, 113:104966, 2022.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Li Hu, Hongyang Yan, Lang Li, Zijie Pan, Xiaozhang Liu, and Zulong Zhang. Mhat: An efficient model-heterogenous aggregation training scheme for federated learning. *Information Sciences*, 560:493–503, 2021.

[18] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33716–33727. Curran Associates, Inc., 2022.

[19] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

[20] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[21] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009.

[22] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

[23] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021.

[24] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[25] Lumin Liu, Jun Zhang, SH Song, and Khaled B Letaief. Communication-efficient federated distillation with active data sampling. In *ICC 2022-IEEE International Conference on Communications*, pages 201–206. IEEE, 2022.

[26] Yong Luo, Liancheng Yin, Wenchao Bai, and Keming Mao. An appraisal of incremental learning methods. *Entropy*, 22(11):1190, 2020.

[27] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3589–3599, 2021.

[28] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[29] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2023.

[30] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[31] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[32] Wenju Sun, Jing Zhang, Danyu Wang, Yangli-ao Geng, and Qingyong Li. Ilcoc: An incremental learning framework based on contrastive one-class classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3580–3588, June 2021.

[33] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.

[34] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[35] Zhiyuan Wu, Sheng Sun, Yuwei Wang, Min Liu, Quyang Pan, Xuefeng Jiang, and Bo Gao. Fedict: Federated multi-task distillation for multi-access edge computing. *IEEE Transactions on Parallel and Distributed Systems*, 2023.

[36] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3014–3023, June 2021.

[37] Guanglei Yang, Enrico Fini, Dan Xu, Paolo Rota, Mingli Ding, Moin Nabi, Xavier Alameda-Pineda, and Elisa Ricci. Uncertainty-aware contrastive distillation for incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2567–2581, 2022.

[38] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019.

[39] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[40] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.