FedLID: Self-Supervised Federated Learning for Leveraging Limited Image Data

Athanasios Psaltis*[1,2], Anestis Kastellos*[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
{kastellosa, daras}@iti.gr {apsaltis, bpatr}@uniwa.gr

Abstract

This study investigates the challenging task of training visual models with very few available data, further complicated by the distribution being imbalanced and scattered across nodes. To address this diverse availability of training data in different federated settings, a customized self-supervised learning approach tailored specifically for each scenario is being proposed. In particular, a hybrid approach combining self-supervised and supervised learning techniques under a federated umbrella has been utilized at both the global and local level, harnessing the potential of unlabeled data. Extensive experiments provide a detailed analysis of the problem at hand and demonstrate the particular characteristics of the proposed learning schemes in distributed scenarios. The overall proposed approach achieves superior recognition performance in the currently broadest public dataset, surpassing all baselines by a substantial margin. The proposed solution can operate efficiently at a local level without prior knowledge of the characteristics or distribution of data across nodes.

1. Introduction

The problem of learning visual models from very few available training data and scattered distribution poses a significant challenge in the field of Machine Learning (ML). The majority of Deep Learning (DL) algorithms in general require a substantial amount of data to achieve the desired performance. However, when it comes to real-world applications, the available data sources are typically limited and often fragmented, making it challenging to apply traditional learning techniques. Many research groups have dedicated resources to find reliable solutions [8], but despite the abundance of published work, the problem of learning meaningful representation from sparse data still poses significant difficulties.

To achieve truly robust recognition performance in various visual analysis tasks where limited data are involved, it is essential to address several key issues. These include ensuring data quality and accurate annotation, mitigating the effects of overfitting and promoting generalization, handling imbalanced class distributions, *etc*. [5]. Initially, the research efforts focused on combinations of data pre-processing techniques, domain adaptation approaches and regularization methods [37, 29, 36, 45, 6, 34, 41, 33]. While these strategies were effective to some extent, recent advances in representation learning techniques [2, 43, 3] have significantly boosted the field, literally transforming the way visual analysis is approached. These techniques enable the extraction of meaningful and discriminative representations from limited data, enhancing the model's generalization power, especially on unseen samples. As a result, the extracted representation can better capture the inherent patterns within the data, enabling models to leverage limited data more effectively. This, in turn, leads to improved performance and robustness in achieving the desired task.

While there has been extensive research on applying representation learning to centralized data settings, the exploration of its application to federated datasets with incomplete annotations and scarce data samples is still relatively limited, due to data privacy concerns, data and system heterogeneity constraints, and limited annotations bottlenecks. Research groups have devoted resources for approaches specifically designed for federated settings, leveraging the power of representation learning while respecting the decentralized nature of data. These approaches often involve a range of learning techniques, such as self-supervised, semi-supervised, unsupervised learning, and transfer learning [44, 39, 26, 18, 11]. In the context of federated learning

*Equal Contribution

(FL), the principal impediment resides in the sensitivity of data, rendering it non-transferable from local users to the central server. Moreover, an associated issue pertains to the quantity of annotated data. The central server may house semi or fully-annotated data, or data that is an amalgamation of different datasets.

In traditional FL a critical limitation faced by local nodes pertains to their constrained resources, both in terms of data processing and storage capabilities. As a result of these constraints, the deployment of both self-supervised and semi-supervised learning methodologies at the local-level becomes infeasible. This challenge accentuates the pressing need for an innovative approach within FL. Such an approach should have the flexibility to alternate between and synergistically combine self-supervised and supervised learning techniques, specifically tailoring them to the distinct dynamics of both local and global FL rounds. The evolution and integration of these methods can potentially optimize the FL process, harnessing the strengths of both self-supervision and traditional supervision within the federated context.

Through the application of self-supervised learning methodologies in our approach, it becomes plausible to surmount the challenges posed by non-annotated data. The proposed specific methodology, we harnessed Constrastive Learning (CLR) [2] with the objective of discerning the optimal representations by drawing contrasts between each instance of the dataset and the remainder thereof. This training modality does not necessitate labels for the data and yields exceptional outcomes, particularly given that the data requisitioned for training do not demand any annotation effort whatsoever.

In a paradigm involving fully-supervised training of the central server, it is incumbent that the data are fully-annotated with a high degree of label quality. However, when employing self-supervised training, there is no requirement for fully annotated data within the central server, hence enabling us to capitalize on all available data situated within the central server to construct a highly potent feature extractor that learns data representations independent of the need for classes.

Contrastingly, local clients acquire representation predicated on the specific annotated dataset that is accessible to them; hence each client learns the optimal representation dedicated to the specific dataset. By possessing a locally fine-tuned model for each distinct local dataset, which when amalgamated on the central server and combined with the powerful feature extractor located on the centralized server, the updated model will embody knowledge from each client and from the expansive dataset in the central server.

The main contributions of this work are summarized as follows:

(a) **A novel self-supervised learning approach that empowers the central server to exploit every piece of data within its possession**, by leveraging CLR techniques to explore meaningful representation and extract informative features at global level even in scenarios where labeled data are limited.

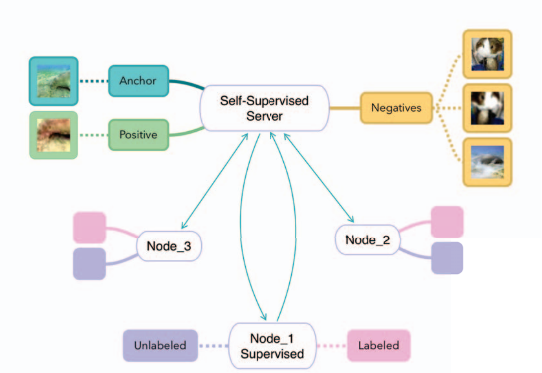

Figure 1: The proposed Federated Learning (FL) scheme utilizes a centralized server to create and share the global model, while several local nodes participate asynchronously in the training process.

(b) **A hybrid FL scheme that seamlessly blends self-supervised and supervised techniques**, adapting them to the unique dynamics of local and global learning rounds within the federated context.

(c) **Validation of the proposed approach by conducting extensive comparisons with a fully-supervised learning process within the same FL scheme**, thereby demonstrating the efficacy of our combination of contrastive and supervised learning on two standard benchmark databases. Extensive experimentation and comparative evaluation highlights the advantages of the proposed schemes under various federated scenarios.

The remainder of this paper is organized as follows: Related work is reviewed in Section 2. The proposed hybrid Federated Self-Supervised 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. Self-supervised Learning

It is undeniable that there are many annotated datasets available nowadays. However, it is clear that we cannot constantly access full databases relevant to every imaginable work. The use of supervised learning approaches, which primarily rely on the availability of properly annotated datasets, is severely constrained by such obstacles. Self-supervised learning techniques, in contrast, have the advantage of not requiring annotations, demonstrating its applicability to a variety of issues. These approaches have shown tremendous promise when applied to semi-annotated or non-annotated data, providing outcomes that are, in fact, notable. They achieve the later by creating a self-supervised job for learning.

Auto-encoders [15] are one such classic example. When compared to the lack of annotated data, they prove to be especially useful. In order to decode the latent representation and return to the original input, autoencoders transpose the input into a latent space representation first. Replicating the input as accurately as possible is the main goal. Despite being an unsupervised scheme, Generative Adversarial Networks (GANs) [9] can function in a self-supervised manner. By learning to perform tasks such as predicting the rotation of an image [4] or filling in missing parts of an image [35], the model discovers rich feature representations of the data without requiring explicit labels, which can then be used for solving demanding computer vision tasks under limited data settings.

Recently, similarity learning techniques have attained particular focus. More specifically, Bootstrap Your Own Latent (BYOL) [10] is a self-supervised learning algorithm that leverages two neural networks, namely a target network and an online network. Instead of relying solely on negative samples for learning, BYOL trains the online network to align its predictions with the output of the target network, which is a slow-moving average of the online network, thereby learning representations from different augmentations of the same image.

2.2. Contrastive Learning

CLR serves as a potent instrument within the realm of self-supervised learning [13, 1, 40]. Its fundamental aim revolves around crafting representations that induce a tendency for similar instances to congregate in the embedding space, while simultaneously propelling dissimilar instances to be dispersed at a greater distance. To fulfill this aspiration, every instance undergoes transformation, yielding a variant distinct from the original yet preserving the underlying principle intact. This is facilitated through the deployment of augmentations that possess the ability to modify the appearance of the instance without distorting its foundational principle [17].

This theory paves the way for a self-supervised learning task, wherein the network endeavors to diminish the distance between the original and the augmented instances in the latent space, while concurrently increasing the space between all the other instances within the dataset. Within this postulate, the network strives to comprehend the representation of the instances in a manner that enhances their separability in the embedding space, thereby simplifying the process of discerning distinct instances.

2.3. Federated Learning

FL is an innovative ML approach that leverages decentralized data and computational resources to deliver more tailored and flexible applications while upholding the privacy of users and organizations. FL has demonstrated exceptional results in numerous visual analysis tasks, such as image classification, object detection and action recognition [25, 12, 32], indicating its robustness and effectiveness in these areas.

The research on FL has focused on increasing communication efficiency and accelerating model updates. McMahan *et al*.'s [27] pioneering work introduced the concept of averaging local stochastic gradient descent updates to increase the calculated quantity of each client between communication rounds. To address low device participation, non-independent and identically distributed (Non-IID) local data, other studies, employ online knowledge distillation approaches, also called codistillation, for communication-efficient FL. Unlike transferring model updates, codistillation focuses on transmitting the local model prediction on a public dataset that is accessible to multiple clients. This method proves beneficial in reducing communication costs, particularly when the size of the local model exceeds the size of the public data[24, 38, 30]. In a recent study [23], researchers introduced a novel approach, named as MOON, where they propose a local model constrastive loss comparing representations of global and local models from successive FL rounds. This technique aims to improve the training of individual parties by conducting CLR in the model-level, specifically in the feature representation space, pushing the current local representation closer to the global representation and further away from the previous local one. Similarly, authors in [28] proposed a distillation- based regularization method, named FedAlign, that promotes local learning generality while maintaining excellent resource efficiency.

In an attempt to leverage the unlabeled data available across multiple nodes, while utilizing the limited labeled samples, several recent studies explored novel algorithms, ranging from self-training, co-training to knowledge distillation. Authors in [16] proposed a novel semi-supervised method, termed as FedMatch, which learns inter-client consistency between nodes, and decomposes model parameters to reduce interference between both supervised and unsupervised tasks. Zhang *et al*. [44] proposed an unsupervised representation learning algorithm, called FedCA, where each client performs unsupervised learning on its local data, leveraging techniques such as CLR to capture meaningful patterns and representations. The learned representations are then aggregated and refined at a central server, resulting in a powerful and comprehensive representation model that encapsulates the knowledge from all distributed sources. In a similar approach, Han *et al*. [11] introduced FedX, an unsupervised FL framework that learns unbiased representation from decentralized and heterogeneous local data, by employing a two-sided knowledge distillation with CLR as a core component. Its adaptable architecture can be used as an add-on module for existing unsupervised algorithms in federated settings. Moreover, two federated self-supervised learning frameworks for images with limited labels was proposed in [39], based on federated CLR with feature sharing (FedCLF). In contrast to previous approaches that assume labeled data are available at the client-side, Long *et al*. [26] introduced FedCon, a novel framework designed to address scenarios where local client data is unlabeled and only the server has access to labeled data. FedCon tackles this challenge by employing a contrastive network architecture, which consists of two subnetworks, enabling effective handling of the unlabeled data at the client-side. Recently, Khowaja *et al*. [18] proposed the SelfFed framework that operates in two distinct phases. The initial phase involves self-supervised pretraining, where a decentralized approach is employed to train a Swin Transformer-based encoder. In the subsequent phase, referred to as fine-tuning, the framework incorporates a contrastive network and introduces a novel aggregation strategy. This phase aims to refine the pre-trained encoder using limited labeled data specific to the target task.

While these approaches have made significant contributions to FL with limited labeled data, they have certain limitations. One limitation is the reliance on labeled data at either the client-side or the server-side. Another limitation is the lack of exploration of meaningful representations and informative features at a global level. In contrast, the proposed method introduces a novel self-supervised learning approach that enables the central server to harness the entirety of available data within its possession. By utilizing CLR techniques, this study explores strong representations and extracts informative features at a global level, even in scenarios where labeled data are limited or unavailable. This allows for more effective utilization of data and enhances the performance and generalization capabilities of the final model.

3. Proposed learning schemes

Problem statement: In a federated system, data are inherently localized to individual clients and the dissemination of this information to other clients is strictly prohibited. On the other hand, the central server, effectively serving as a simulated environment for the local clients, has the ability to leverage a substantial volume of data for training purposes, thereby accommodating a broad array of data variations. The goal is to train each local model separately to its own subset of data, while the centralized server is trained in a self-supervised manner on vast database, with high diversity, regardless whether the data are annotated or not, whereas the complete system of the local clients maintain uniformity in the overall accuracy ascertained in the common test set.

In the methodology that we advocate, a hybrid learning scheme involving self-supervised and supervised training strategies is employed. Owing to the substantial magnitude of images contained within the Tiny-ImageNet[21] dataset, it serves as an ideal candidate for the training of the central server through the implementation of unsupervised CLR. On the other hand, for the cultivation of individual local clients, the CIFAR-100[20] dataset is deployed, partitioned in both a balanced and unbalanced manner to facilitate experiments that cater to both IID and Non-IID conditions. CIFAR-100 and Tiny-ImageNet are used since they are contextually similar as shown in figure 2.

3.1. FedLID Local Supervision

The local clients are trained in a supervised fashion utilizing the CIFAR-100 dataset. A distinct segment of the dataset is allocated for each client that remains consistent throughout all federated rounds, on which they receive training. The training portion of the dataset is utilized for the purpose of training, while the validation subset serves to authenticate each of the clients' models. Notably, the validation set is communal, hence accessible to all clients. For the purpose of training the clients' networks, a classification head was appended that maps the feature vectors to the total number of classes present in the dataset, amounting to 100. Cross Entropy was deployed as the loss function for the training process, the formula is displayed below:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{C} y_{ij}\log p_{ij} \tag{1}$$

where N is the total number of images, C the total number of classes, y_{ij} the label of the i, j image and p_{ij} the probability of the class. Adam[19] algorithm was chosen as the optimization strategy.

3.2. FedLID Global Supervision

In the federated 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. Each image from the said dataset serves as the anchor image x_m^i. To generate a slightly variant, positive image denoted x_p^i, image augmentations are utilized. Both 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[2].

The embedding dimensions of the feature map encompass 126 channels. For the purpose of projection, the final fully connected layer of the model is omitted, and a novel projection head is affixed that maps the 512-dimensional latent space to a 126-dimensional space employing a Rectified Linear Unit (ReLU)[31] activation function. The feature maps of both the anchor image $\Phi_m^i = z_i$ and the positive image $\Phi_p^i = z_j$ are harnessed for the calculation of the contrastive loss. The formula of the loss is indicated below:

$$\mathcal{L}_{i,j} = \log \frac{\exp\left(sim(z_i, z_j)/\mathcal{T}\right)}{\sum_{k=1,[k\neq i]}^{2N} \exp\left(sim(z_i, z_k/\mathcal{T}\right)} \tag{2}$$

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. Through the implementation of contrastive loss, the network strives to learn representations that induce a proximity between the anchor image and the positive image in the embedding space, whilst ensuring a separation between the anchor image and all other images within the dataset. 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 class.

3.3. Federated aggregation

Optimization strategies and in particular aggregation algorithms play an important role in FL as they are responsible for combining the knowledge from all devices/nodes while respecting data's privacy. Although adapting FL to fully-supervised federated schemes seems to be straightforward, shifting to more complex schemes that involve further self-supervision steps, as the ones mentioned earlier, can prove to be more challenging than anticipated. Prior to the application of the proposed self-supervision step, federated optimization is realized for the locally fully-supervised trained models (Figure 1). In particular, the broadest aggregation mechanism, namely FedAvg, has been followed, examined in detail and adapted to the specific scenarios. Firstly, the server distributes the initial version of the model to each node for training on local data. This first version can either be pre-trained on a predefined dataset (*i.e.* ImageNet1k[7]), or be initialized randomly. In each round, the algorithm performs a set of local model updates (*i.e.* cross-entropy loss) on a subset of nodes, followed by a server-side aggregation task, trying to minimize the following objective function, which is actually the sum of the weighted average of the clients' local errors, where F_k is the local objective function for the kth device and p_k specifies the relative impact of each device:

$$\min_{w} = \sum_{k=1}^{N} p_k F_k(w) \tag{3}$$

In the final step, the updated global model is forwarded to the local nodes for another round of training. The process is continuing until the global aggregated model is fully trained and achieves the desired performance.

3.4. FedLID Algorithm

In this section, the complete algorithm of the described procedure is presented. The entire process is divided into two main algorithms, the first one describes the required steps for training and updating the network in the main server, while the second one defines the action for training and updating the network in each local client.

Algorithm 1 FedLID Algorithm

```
Require: $\mathcal{C}$ is the total number of clients, $\theta_g$ represents the parameters of the global model, $\theta_l^i$ represents the parameters of the local models, $\mathcal{D}_l^i$ are the separate datasets for the local clients, $\mathcal{D}_g$ is the dataset of the central server, $\mathcal{D}_val$ is the common validation dataset, and $\eta$ is the learning rate.
Initialize global model $\theta_g$
for each $i$ from 1 to $C$ do
    Initialize local models $\theta_l^i$
end for
Prepare global dataset $D_g$
for each $i$ from 1 to $C$ do
    Prepare local datasets $D_l^i$
end for
Prepare common local dataset for validation $D_{val}$
Server executes:
for each Federated round $t = 0, 1, 2, \ldots$ do
    for each client in parallel do
        $\theta_{l_{t+1}} \leftarrow$ ClientUpdate$(D_l, \theta_{l_t})$
    end for
    $\theta_g \leftarrow$ FedAvg$(\theta_l) Eq.3$
    for each global epoch $i$ from 1 to $G$ do
        for each batch in $D_g$ do
            $\theta_g \leftarrow \theta_g - \eta \nabla L(\theta_g)$    ▷ Update the global model with Eq. 2 (Contrastive Loss)
        end for
    end for
    Distribute the updated global model to the clients
    Validate the updated models on $D_val$
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 $\theta_l$
```

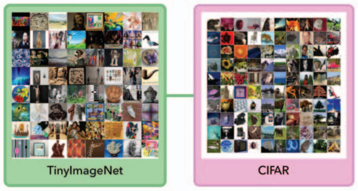

Figure 2: Images from Tiny-ImageNet (left) and the CIFAR-100 (right), showcasing the contextual similarities of the two datasets.

4. Experimental results

4.1. Data settings

Within this segment, experimental results from the application of the proposed Federated Self-Supervised learning strategies are presented. The evaluation process utilized the CIFAR-100 image classification dataset as a benchmark. This dataset was adapted into both IID and Non-IID variations to reflect different FL scenarios. One is replicating the balanced data, *i.e*. IID dataset, while the other is the extreme example of a Non-IID dataset. From both datasets, a set of 100 categories was used and with the total number of participating nodes being 10, among which the images are to be distributed.

4.1.1 IID and Non-IID Settings

Like previous studies [42], the parameter $\alpha > 0$ controls the identicalness among participants. Different α values were tested, where with $\alpha \rightarrow \infty$, all participants have identical distributions and $\alpha \rightarrow 0$, each participant has examples from only one class. To support IID setting, the dataset was divided with medium heterogeneity by setting $\alpha = 1$. Therefore, a node can have images from any number of classes. In contrast, for the case of the Non-IID set, the original CIFAR-100 dataset was divided, with higher level of heterogeneity by setting $\alpha = 0.5$. Here, nodes tend to have significant number of samples from some classes and few or no samples for the other classes. Each node randomly sampled $\frac{1}{10}$ of training and validation data respectively, while the test set data were left out for the final system evaluation, both at global and local level.

4.1.2 Image augmentations

For the CLR scheme, a data augmentation strategy as the one in SimCLR[2] is adopted. Initially, a stochastic crop of the image is procured, which is subsequently subjected to a random horizontal flip. This is then followed by arbitrary distortion of brightness, contrast, hue, and saturation parameters, complemented by an optional grayscale transformation. Subsequently, a Gaussian blur filter is administered as the terminal step of the augmentation process. The image is ultimately resized to dimensions of 224 × 224 and undergoes normalization.

4.2. Implementation details

4.2.1 Architecture and Parameters

The architecture of the proposed method employs a distributed framework, composed of a Convolutional Neural Network (CNN)[22], specifically ResNet18[14] is used for experiments because it is shallow, fast in training and has satisfactory results, that is deployed both at the centralized server and the local clients, albeit with unique modifications for each entity. The step-by-step process of the training of the federated scheme is illustrated in Algorithm 1. For the central server, the final fully connected layer of the network is replaced with a projection head, enabling the transition of the ResNet feature map to a new embedding space, thereby facilitating the CLR scheme. Conversely, the local clients are trained via supervision, resulting in the substitution of the final fully connected layer with a classification head that maps the feature vectors to the corresponding classes, the full algorithm of the local supervision is depicted in 2. Regarding the optimizer, Adam is employed with a learning rate of 0.001 and the batch size of the system is 256. It is crucial to note that the ResNet model retains knowledge acquired from the ImageNet1k[7], as it operates on the basis of pre-trained weights. To facilitate an unbiased comparison between the proposed approach and leading-edge federated systems, additional experiments are conducted employing the ResNet50 architecture.

4.2.2 Federated setting

In the context of the training paradigm for the suggested FL system – a system comprising both a centralized server and 10 local clients – a comprehensive training period of 10 federated rounds is undertaken. Within each of these rounds, the local clients individually embark on a training process spanning 10 epochs, operating in isolation without any inter-client communication. Upon completion of their respective training, the local clients communicate with the central server, transferring their uniquely derived weights. This information is assimilated by the central server which proceeds to aggregate the weights and undergoes a training cycle for 60 epochs prior to disseminating the updated model to the local clients. Post distribution, the refreshed model undergoes an evaluation process leveraging the common test set accessible to all clients.

4.2.3 Baselines

To affirm the validity of FedLID as unequivocally as possible, a comparative performance analysis across two distinct configurations was executed. Initially, benchmark experiments were undertaken within a fully-supervised federated framework, in both IID and Non-IID scenarios, wherein the volume of data accessible on the central server progressively reduced in increments of 20%, ranging from 100% to 20%. Subsequently, a comparative evaluation was conducted against the following cutting-edge benchmarks: FedCA and FedSimCLR [44].

4.2.4 Training Setup

The proposed federated system's training was executed on a single computer, outfitted with a GeForce RTX 3090 (VRAM 24 Gbs) and furnished with 32 Gbs of RAM.

4.3. Performance Evaluation

4.3.1 IID Setting

The presentation of results, under circumstances of equitably distributed data amongst the local clients, is delineated below. As discernible from Table 1, when juxtaposed with the fully-supervised framework, FedLID provides superior results (47.52%) with relative improvement over the baseline of 8.05%, even when pitted against the optimum case scenario of supervision (43.98%), which implies the utilization of 100% of the data situated within the central server. As anticipated, the overall accuracy experiences a decline with a diminishing quantity of training data in the server. Consequently, in situations where the central server houses extremely limited annotated data, the proposed method surpasses the conventional supervised FL. Figure 3 shows the impact that the training on the server has in the overall accuracy of a client. The underlying cause for the observed outcome is that each local client independently assimilates the representation of its specific training subset. However, when these learned representations converge on the central server, this disparate information is consolidated. Coupled with the CLR setting deployed on the server, these combined representations converge closer to the authentic data distribution of the test set.

Percentage of Data in the Central Server	Average Accuracy over the Local Clients
20%	42.15%
40%	41.79%
60%	42.74%
80%	43.81%
100%	43.98%
FedLID(Ours)	**47.52%**

Table 1: Comparison of our method with different percentages of available data in the central server. The data in the local clients are divided equally.

It is subsequently discerned that FedLID surpasses the performance metrics of contemporary state-of-the-art algorithms as shown in Table 2. This superior performance of 47.52% is achieved even when utilizing shallower network architectures and a larger number of clients, both factors contributing to reduced training data for each individual local client.

Method	Architectute	Clients	CIFAR100
FedSimCLR	ResNet50	5	34.18%
FedCA	ResNet50	5	39.47%
FedLID(Ours)	ResNet18	10	**47.52%**

Table 2: Average Accuracy over the local clients on CIFAR-100 udner the IID setting with $\alpha = 1$.

4.3.2 Non-IID Setting

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, FedLID achieves superior results (48.88%), as depicted in Table 3, by leveraging the inherent structure and relationships within the global data, with relative improvement of 39.02% over the baseline system. Figure 4 demonstrates the effect that the main server has in performance of the client. By pretraining the global aggregated model on a self-supervised task and subsequently fine-tuning it on the target local supervised task, the model can effectively exploit the knowledge gained at global level while adapting to the specific characteristics of the Non-IID dataset, resulting in improved accuracy and generalization at local level. It is entirely plausible in practical federated systems that the data distributed among local clients may not be evenly distributed, and the server data could be semi-annotated or non-annotated. Under such realistic conditions, the proposed methodology outperforms all baseline cases.

Percentage of Data in the Central Server	Average Accuracy over the Local Clients
20%	32.83%
40%	33.08%
60%	32.84%
80%	34.32%
100%	35.16%
FedLID(Ours)	**48.88%**

Table 3: Comparison of our method with different percentages of available data in the central server. The data in the local clients are divided in an imbalanced way.

In relation to a comparison with state-of-the-art methods, FedLID demonstrates exceptional performance, outpacing the next closest approach by a 10% margin, as portayed in Table 4. This impressive outcome is achieved despite the fact that each client has less data, and the network used is smaller in terms of parameters.

It is evident from the results that the proposed method exhibits improved accuracy on Non-IID data compared to IID data distributions (48.88% and 47.52% respectively). This is due to the fact that in the IID setting, the training data across nodes is representative of the overall population, and the local model can learn common patterns more

Method	Architectute	Clients	CIFAR100
FedSimCLR	ResNet50	5	33.63%
FedCA	ResNet50	5	38.94%
FedLID (Ours)	ResNet18	10	**48.88%**

Table 4: Average Accuracy over the local clients on CIFAR-100 udner the Non-IID setting with $\alpha = 0.5$.

effectively, leading to fast convergence. In addition, the pretext task used for self-supervision is almost aligned with the target supervised task, thus the fine-tuning process on the balanced (*i.e.* uniformly distributed accross nodes) dataset does not struggle to align the learned representation with the task-specific requirements; hence leading to less benefit from self-supervision step. On the contrary, self-supervised methods often excel in scenarios where there are variations, complex dependencies, or imbalances in the data distribution, as they can capture the underlying structure and extract meaningful representations. Consequently, in homogeneous scenarios, the impact of self-supervision may be relatively diminished, thus largely explains the difference in performance.

4.3.3 Weakly annotated setting

In an effort to investigate the impact on performance, we intentionally reduced the available data to 10%, aiming to create conditions of partial annotation in the local nodes. This approach allows us to study the consequences both prior to and following the application of the proposed self-supervised method at global level. The results demonstrate that despite the significant reduction in local data and a limited number of training iterations (10 local epochs and 10 federated rounds), the performance slightly increases compared to the baseline approach (*i.e.* FedCA) in the IID setting while the comparison is favorable in the context of Non-IID, as depicted in Table 5. It is evident that the limited number of local epochs hampers performance improvement, and the local model has potential for further convergence, as illustrated by the accompanying diagrams. Furthermore, it can be observed that increasing the number of communication rounds consistently enhances performance, showcasing the cumulative strength of the self-supervised model in the federated aggregation process.

Label Fraction	Setting	Method	CIFAR100
10%	IID	FedCA	32.09%
		FedLID(Ours)	**33.51%**
	Non-IID	FedCA	22.46%
		FedLID(Ours)	**24.12%**

Table 5: Accuracy of the state-of-the-art in weakly annotated scheme with different label ratios. FedCA's total number of local clients is *5*, while FedLID is utilizing a federated system with *10* clients in total.

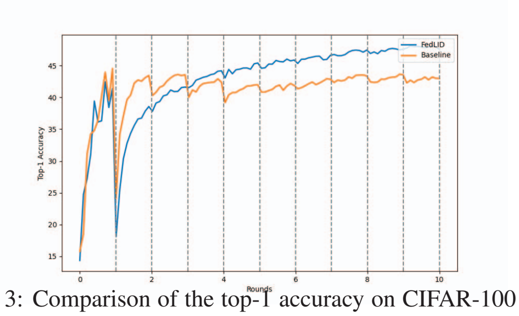

Figure 3: Comparison of the top-1 accuracy on CIFAR-100 in an IID setting, over the federated rounds of FedLID and the baseline. The dotted line is the aggregation of the local models and the trainin

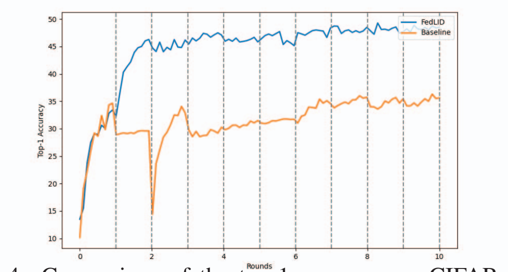

Figure 4: Comparison of the top-1 accuracy on CIFAR-100 in a Non-IID setting, over the federated rounds of FedLID and the baseline. The dotted line is the aggregation of the local models and the training of the centralized server.

5. Conclusions

This study addresses the challenges of training visual models with limited and scattered data by leveraging the power of representation learning techniques under the FL paradigm. It explores various methodologies, including self-supervised, and unsupervised learning, in the context of federated settings. The proposed approach empowers the central server to leverage every piece of data within its possession, even in scenarios where labeled data is limited or non-existent. By utilizing CLR techniques, the approach explores meaningful representations and extracts informative features at a global level. The proposed solution can operate efficiently at a local level without prior knowledge of the characteristics or distribution of data across nodes. Additionally, it offers substantial improvements, particularly in cases where data are significantly limited. The extensive experimentation and comparative evaluation validate the effectiveness of the proposed approach, highlighting its advantages over traditional fully-supervised learning methods. This research contributes to the advancement of self-supervised learning in federated settings, offering promising opportunities for robust visual analysis tasks with limited data.

Acknowledgment

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

References

[1] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

[2] 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.

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[4] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12154–12163, 2019.

[5] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *Proceedings of the European conference on computer vision (ECCV)*, pages 268–283, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[11] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. Fedx: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*, pages 691–707. Springer, 2022.

[12] Chaoyang He, Alay Dilipbhai Shah, Zhenheng Tang, Di Fan1Adarshan Naiynar Sivashunmugam, Keerti Bhogaraju, Mita Shimpi, Li Shen, Xiaowen Chu, Mahdi Soltanolkotabi, and Salman Avestimehr. Fedcv: a federated learning framework for diverse computer vision tasks. *arXiv preprint arXiv:2111.11066*, 2021.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[14] 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.

[15] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[16] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*, 2020.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[18] Sunder Ali Khowaja, Kapal Dev, Syed Muhammad Anwar, and Marius George Linguraru. Selffed: Self-supervised federated learning for data heterogeneity and label scarcity in iomt. *arXiv preprint arXiv:2307.01514*, 2023.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[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] 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.

[25] Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13172–13179, 2020.

[26] Zewei Long, Jiaqi Wang, Yaqing Wang, Houping Xiao, and Fenglong Ma. Fedcon: A contrastive framework for federated semi-supervised learning. *arXiv preprint arXiv:2109.04533*, 2021.

[27] 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.

[28] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022.

[29] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[30] Zijia Mo, Zhipeng Gao, Chen Zhao, and Yijing Lin. Feddq: A communication-efficient federated learning approach for internet of vehicles. *Journal of Systems Architecture*, 131:102690, 2022.

[31] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[32] Athanasios Psaltis, Charalampos Z. Patrikakis, and Petros Daras. Deep multi-modal representation schemes for federated 3d human action recognition. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, page 334–352, Berlin, Heidelberg, 2023. Springer-Verlag.

[33] Guo-Jun Qi, Liheng Zhang, Hao Hu, Marzieh Edraki, Jingdong Wang, and Xian-Sheng Hua. Global versus localized generative adversarial nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1517–1525, 2018.

[34] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018.

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[36] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.

[37] Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. Enaet: A self-trained framework for semi-supervised and supervised learning with ensemble transformations. *IEEE Transactions on Image Processing*, 30:1639–1647, 2020.

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

[39] Yawen Wu, Dewen Zeng, Zhepeng Wang, Yi Sheng, Lei Yang, Alaina J James, Yiyu Shi, and Jingtong Hu. Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis. *arXiv preprint arXiv:2208.11278*, 2022.

[40] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[41] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[42] 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.

[43] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1476–1485, 2019.

[44] Fengda Zhang, Kun Kuang, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Yueting Zhuang, and Xiaolin Li. Federated unsupervised representation learning. *arXiv preprint arXiv:2010.08982*, 2020.

[45] Liheng Zhang and Guo-Jun Qi. Wcp: Worst-case perturbations for semi-supervised deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3912–3921, 2020.