Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Federated unsupervised representation learning*

Fengda ZHANG†1, Kun KUANG†‡1, Long CHEN1, Zhaoyang YOU1, Tao SHEN1, Jun XIAO1, Yin ZHANG1, Chao WU2, Fei WU1, Yueting ZHUANG1, Xiaolin LI3,4,5

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*School of Public Affairs, Zhejiang University, Hangzhou 310027, China*

[3]*Tongdun Technology, Hangzhou 310000, China*

[4]*Institute of Basic Medicine and Cancer, Chinese Academy of Sciences, Hangzhou 310018, China*

[5]*ElasticMind.AI Technology Inc., Hangzhou 310018, China*

†E-mail: fdzhang@zju.edu.cn; kunkuang@zju.edu.cn

Received June 21, 2022; Revision accepted Oct. 27, 2022; Crosschecked July 30, 2023

Abstract: To leverage the enormous amount of unlabeled data on distributed edge devices, we formulate a new problem in federated learning called federated unsupervised representation learning (FURL) to learn a common representation model without supervision while preserving data privacy. FURL poses two new challenges: (1) data distribution shift (non-independent and identically distributed, non-IID) among clients would make local models focus on different categories, leading to the inconsistency of representation spaces; (2) without unified information among the clients in FURL, the representations across clients would be misaligned. To address these challenges, we propose the federated contrastive averaging with dictionary and alignment (FedCA) algorithm. FedCA is composed of two key modules: a dictionary module to aggregate the representations of samples from each client which can be shared with all clients for consistency of representation space and an alignment module to align the representation of each client on a base model trained on public data. We adopt the contrastive approach for local model training. Through extensive experiments with three evaluation protocols in IID and non-IID settings, we demonstrate that FedCA outperforms all baselines with significant margins.

Key words: Federated learning; Unsupervised learning; Representation learning; Contrastive learning
https://doi.org/10.1631/FITEE.2200268 **CLC number:** TP183

1 Introduction

Federated learning (FL) is proposed as a paradigm that enables distributed clients to collaboratively train a shared model while preserving data privacy (McMahan et al., 2017). Specifically, in each round of FL, clients obtain the global model and update it on their own private data to generate the local models, and then the central server aggregates these local models into a new global model. Most of existing works focus on supervised FL, in which clients train their local models with supervision. However, the data generated in edge devices are typically unlabeled. Therefore, learning a common representation model for various downstream tasks from decentralized and unlabeled data while keeping private data on devices, i.e., federated unsupervised representation learning (FURL), remains still an open problem.

It is a natural idea that we can combine FL with unsupervised approaches, which means that clients can train their local models via unsupervised methods. There are a lot of highly successful works on

‡ Corresponding author

* Project supported by the National Key Research & Development Project of China (Nos. 2021ZD0110700 and 2021ZD0110400), the National Natural Science Foundation of China (Nos. U20A20387, U19B2043, 61976185, and U19B2042), the Zhejiang Natural Science Foundation, China (No. LR19F020002), the Zhejiang Innovation Foundation, China (No. 2019R52002), and the Fundamental Research Funds for the Central Universities, China

ORCID: Fengda ZHANG, https://orcid.org/0000-0001-5280-413X; Kun KUANG, https://orcid.org/0000-0001-7024-9790

unsupervised representation learning. Particularly, contrastive learning methods train models by reducing the distance between representations of positive pairs (e.g., different augmented views of the same image) and increasing the distance between negative pairs (e.g., augmented views from different images), which have been outstandingly successful in practice (van den Oord et al., 2019; Chen T et al., 2020; Chen XL et al., 2020; He KM et al., 2020). However, their successes highly rely on their abundant data for representation training; for example, contrastive learning methods need a large number of negative samples for training (Sohn, 2016; Chen T et al., 2020). Moreover, few of these unsupervised methods take the problem of data distribution shift into account, which is a common practical problem in FL. Hence, it is difficult to combine FL with unsupervised approaches for FURL.

In FL applications, however, the data collected by each client are limited and the data distribution of the client might be different from each other (Jeong et al., 2018; Yang Q et al., 2019; Sattler et al., 2020; Kairouz et al., 2021; Zhao et al., 2022). Hence, we face the following challenges in combining FL with unsupervised approaches for FURL:

1. Inconsistency of representation spaces

In FL, the limited data of each client would lead to variation of data distribution from client to client, resulting in inconsistency of representation spaces encoded by different local models (Kuang et al., 2020). For example, as shown in Fig. 1a, client 1 has only images of cats and dogs, and client 2 is with only images of cars and planes. Then, the locally trained model on client 1 encodes only a feature space of cats and dogs, failing to map cars or planes to the appropriate representations, and the same goes for the model trained on client 2. Intuitively, the performance of the global model aggregated by these inconsistent local models may fall short of expectations.

2. Misalignment of representations

Even if the training data of the clients are independent and identically distributed (IID) and the representation spaces encoded by different local models are consistent, there may be misalignment between representations due to randomness in the training process. For instance, for a given input set, the representations generated by a model are equivalent to the representations generated by another model when rotated by a certain angle, as shown in Fig. 1b. It should be noted that the misalignment between local models may have drastic detrimental effects on the performance of the aggregated model.

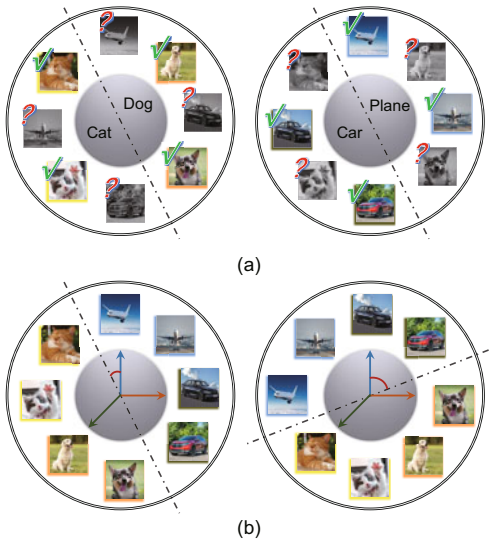

Fig. 1 Illustration of challenges in federated unsupervised representation learning (FURL): (a) inconsistency of representation spaces (data distribution shift among clients causes local models to focus on different categories); (b) misalignment of representations (without unified information, the representation across clients would be misalignment, e.g., rotated by a certain angle). The hyperspheres are representation spaces encoded by different local models in federated learning (FL)

To address these challenges, we propose a contrastive loss-based FURL algorithm called the federated contrastive averaging with dictionary and alignment (FedCA), which consists of two main novel modules: a dictionary module for addressing the inconsistency of representation spaces and an alignment module for aligning the representations across clients. Specifically, the dictionary module, which is maintained by the server, aggregates the abundant representations of samples from clients and these can be shared with each client for local model optimization. In the alignment module, we first train a base model based on small public data (e.g., a subset of STL-10 dataset) (Coates et al., 2011) and then require all local models to mimic the base model such that the representations generated by different local models can be aligned. Overall, in each round, FedCA involves two stages: (1) clients train local representation models on their own unlabeled data via contrastive learning with the two modules mentioned above, and then generate local dictionaries,

and (2) the server aggregates the trained local models to obtain a shared global model and integrates the local dictionaries into a global dictionary.

To the best of our knowledge, FedCA is the first algorithm designed for the FURL problem. Our experimental results show that FedCA has better performance than those naive methods that solely combine FL with unsupervised approaches. We believe that FedCA will serve as a critical foundation in this novel and challenging problem.

2 Related works

2.1 Federated learning

FL enables distributed clients to train a shared model collaboratively while keeping private data on devices (McMahan et al., 2017). Li T et al. (2020) added a proximal term to the loss function to keep local models close to the global model. Wang HY et al. (2020) proposed a layer-wise FL algorithm to deal with the permutation invariance of neural network parameters. However, existing works focus only on the consistency of parameters, while we emphasize the consistency of representations in this study. Some works also focus on reducing the communication of FL (Konečný et al., 2017). To further protect the data privacy of clients, cryptography technologies have been applied to FL (Bonawitz et al., 2017).

2.2 Unsupervised representation learning

Learning high-quality representations is important and essential for various downstream tasks (Zhou et al., 2017; Duan et al., 2018). There are two main types of unsupervised representation learning methods: generative and discriminative (Zhuang YT et al., 2017; Lei et al., 2020; Zhu et al., 2020). Generative approaches learn representations by generating pixels in the input space (Hinton and Salakhutdinov, 2006; Kingma and Welling, 2014; Radford et al., 2016). Discriminative approaches train a representation model by performing pretext tasks, where labels are generated for free from unlabeled data (Pathak et al., 2017; Gidaris et al., 2018). Among them, contrastive learning methods achieve excellent performance (van den Oord et al., 2019; Chen T et al., 2020; Chen XL et al., 2020; He KM et al., 2020). The contrastive loss was proposed by Hadsell et al. (2006). Wu ZR et al. (2018) proposed an unsupervised contrastive learning approach based on a memory bank to learn visual representations. Wang TZ and Isola (2020) pointed out two key properties, namely, closeness and uniformity, related to the contrastive loss. Other works also applied contrastive learning to videos (Sermanet et al., 2018; Tian et al., 2020), natural language processing (NLP) (Mikolov et al., 2013; Logeswaran and Lee, 2018; Yang ZL et al., 2019), audios (Baevski et al., 2020), and graphs (Hassani and Ahmadi, 2020; Qiu et al., 2020).

2.3 Federated unsupervised learning

Before the FL was proposed, there have been some works on unsupervised representation learning in the distributed/decentralized setting, which are easily portable to the FL setting (Kempe and McSherry, 2008; Liang et al., 2014; Shakeri et al., 2014; Raja and Bajwa, 2016; Wu SX et al., 2018). However, different from the deep learning method, the convergence of these methods is limited by the size of the data, and it is difficult to achieve good performance on downstream tasks (Lyu, 2020; Pan, 2020; Zhuang YT et al., 2020).

Some concurrent works (van Berlo et al., 2020; Jin et al., 2020) also focus on FL from unlabeled data with the deep learning method. Different from these works which simply combine FL with unsupervised approaches, we explore and identify the main challenges in FURL and design an algorithm to deal with these challenges. There are some later works aiming to solve our proposed problem (Zhuang WM et al., 2021b). For example, Sattler et al. (2021) proposed to use the unlabeled auxiliary data in FL by federated distillation techniques.

2.4 Contrastive learning for FL

To our best knowledge, our work is the first one to combine contrastive learning with FL, which has inspired some later works (He CY et al., 2021; Ji et al., 2021; Shi et al., 2022). Li QB et al. (2021) conducted contrastive learning at the model level to correct local training. Wu YW et al. (2021) proposed to exchange the features of clients to provide diverse contrastive data to each client. Zhuang WM et al. (2021a) focused on unsupervised setting in FL by designing a dynamically contrastive module with an effective communication protocol. Zhuang WM et al. (2022) proposed a new method to tackle the non-IID

data problem in FL and filled in the gap between FL and self-supervised approaches based on Siamese networks.

3 Preliminaries

In this section, we discuss the primitives needed for our approach. The symbols and the corresponding meanings are given in Table 1.

3.1 Federated learning

In FL, each client $u \in U$ has a private dataset D^u of training samples, and our aim is to train a shared model while keeping private data on devices. There are a lot of algorithms designed for aggregation in FL (Li T et al., 2020; Wang HY et al., 2020), and we point out that our approach does not depend on the way of aggregation. Here, for simplicity, we introduce a standard and popular aggregation method named FedAvg (McMahan et al., 2017). In round t of FedAvg, the server randomly selects a subset of clients $U_t \subseteq U$ and each client $u \in U_t$ locally updates the global model with parameters θ_t on dataset D^u via the stochastic gradient descent rule to generate the local model:

$$\theta_{t+1}^u \leftarrow \theta_t - \eta \bigtriangledown \mathcal{L}(D^u, \theta_t), \tag{1}$$

where η is the stepsize and $\mathcal{L}(D^u, \theta_t)$ is the loss function of client u in round t. Then the server gathers the parameters of the local models $\{\theta_{t+1}^u | u \in U_t\}$ and aggregates these local models via weighted average to generate a new global model:

$$\theta_{t+1} \leftarrow \sum_{u \in U_t} \frac{|D^u|}{\sum_{i \in U_t} |D^i|} \theta_{t+1}^u. \tag{2}$$

The training process above is repeated until the global model converges.

Table 1 Symbols and the corresponding meanings

Symbol	Meaning
U	Set of client indexes $\{1, 2, ..., n\}$
D^u	Local dataset of client u
D_{align}	Additional dataset for alignment
U_t	Set of selected client indexes in round t
C	Proportion of selected clients per round
θ_t	Global model parameters in round t
θ_t^u	Local model parameters of client u in round t
θ_{align}	Alignment model parameters
η	Learning rate for model update
E	Number of local epochs
dict_t	Global dictionary in round t
dict_t^u	Local dictionary of client u in round t
β	Weight of alignment loss
$\boldsymbol{x}$	Input sample
$\boldsymbol{v}$	View augmented from a sample
$f(\cdot)$	Encoder
$\boldsymbol{h}$	Representation for downstream tasks
$g(\cdot)$	Projection head
$\boldsymbol{z}$	Latent projections for contrastive loss
α	Momentum parameter for dictionary update
$\tilde{\boldsymbol{z}}$	Normalized projections
dict	Global dictionary
$\boldsymbol{Z}$	Projection in local ensemble dictionary
dict^u	Local dictionary of client u

3.2 Unsupervised contrastive learning

Unsupervised contrastive representation learning methods learn representations from unlabeled data by reducing the distance between representations of positive samples and increasing the distance between representations of negative samples. Among them, SimCLR achieves outstanding performance and can be applied to FL easily (Chen T et al., 2020). SimCLR randomly samples a minibatch of N samples and executes twice random data augmentations for each sample to obtain $2N$ views. Typically, the views augmented from the same image are treated as positive samples and the views augmented from different images are treated as negative samples (Dosovitskiy et al., 2014). The loss function for a positive pair of samples (i, j) is defined as follows:

$$l_{i,j} = -\ln \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}, \tag{3}$$

where τ is the temperature and $\mathbb{1}_{[k \neq i]} = 1$ if and only if $k \neq i$. $\text{sim}(\cdot, \cdot)$ measures the similarity of two representations of samples (e.g., cosine similarity). The model (consisting of a base encoder network f to extract representation $\boldsymbol{h}$ from augmented views and a projection head g to map representation $\boldsymbol{h}$ to $\boldsymbol{z}$) is trained by minimizing the loss function above. Finally, we use representation $\boldsymbol{h}$ to perform downstream tasks.

4 Method

In this section, we analyze the two challenges mentioned above and detail the dictionary module and alignment module designed for these challenges. Then we introduce the FedCA algorithm for FURL.

4.1 Dictionary module for inconsistency challenge

FURL aims to learn a shared model that maps data to representation vectors such that similar samples are mapped to nearby points in the representation space so that the features are well clustered by classes. However, the presence of non-IID data presents a great challenge to FURL. Since the local dataset D^u of a given client u likely contains samples of only a few classes, the local models may encode inconsistent spaces, causing bad effects on the performance of the aggregated model.

To empirically verify this, we visualize the representations of images from CIFAR-10 via the t-distributed stochastic neighbor embedding (T-SNE) method. To be specific, we split the training data of CIFAR-10 into five non-IID sets, and each set consists of 10 000 samples from two classes. Then, the FedAvg algorithm is combined solely with the unsupervised approach (SimCLR) to learn representations from these sets. We use the local model in the 20^{th} round of the client who has only samples of class 0 and class 1 to extract features from the test set of CIFAR-10 and visualize the representations after dimensionality reduction by T-SNE (Fig. 2a). We find that the scattered representations of samples from class 0 and class 1 are spread over a very large area of representation space, and it is difficult to distinguish samples of class 0 and class 1 from others. It suggests that the local model encodes a representation space of samples of class 0 and class 1, and it cannot map samples of other classes to the suitable positions. The visualization results support our hypothesis that the representation spaces encoded by different local models are inconsistent in a non-IID setting.

We argue that the cause of inconsistency is that the clients can use only their own data to train the local models but the distribution of data varies from client to client. To address this issue, we design a dictionary module (Fig. 3b). Specifically, in each communication round, clients use the global model (including the encoder and the projection head) to obtain the normalized projections $\{\tilde{\boldsymbol{z}}_i\}$ of their own samples and send the normalized projections to the server along with the trained local models. Then, the server gathers the normalized projections into a shared dictionary. For each client, the global dictionary dict with K projections is treated as a normalized projection set of negative samples for local contrastive learning. Specifically, in the local training process, for a given minibatch $\boldsymbol{x}_{\text{batch}}$ with N samples, we randomly augment them to obtain $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and generate normalized projections $\tilde{\boldsymbol{z}}_i$ and $\tilde{\boldsymbol{z}}_j$. Then we calculate the following:

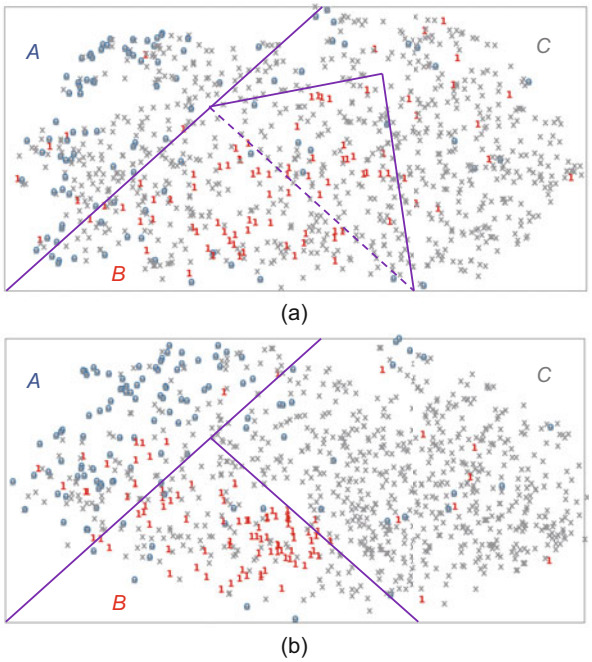

Fig. 2 ***t*-distributed stochastic neighbor embedding (T-SNE) visualization results of representations on CIFAR-10. In federated learning (FL) with a non-independent and identically distributed (non-IID) setting, we use the local model of the client who has only samples of class 0 and class 1 to generate representations. We compare two methods: (a) vanilla federated unsupervised approach FedSimCLR (SimCLR is combined with FedAvg directly) and (b) FedCA (ours). *A* and *B* are the regions where representations of samples of class 0 and class 1 cluster, respectively, and *C* is the remaining region**

$$\text{logits}_{\text{batch}} = \tilde{\boldsymbol{z}}_i \cdot \tilde{\boldsymbol{z}}_j^{\text{T}}, \tag{4}$$

$$\text{logits}_{\text{dict}} = \tilde{\boldsymbol{z}}_i \cdot \text{dict}^{\text{T}}, \tag{5}$$

$$\textbf{logits} = \text{concat}\big([\text{logits}_{\text{batch}}, \text{logits}_{\text{dict}}], \text{dim} = 1\big), \tag{6}$$

where concat() denotes concatenation, the size of **logits** is $N \times (N + K)$, and dim=1 means that they are concatenated in the 1^{st} dimension. Now, we turn the unsupervised problem into an $(N + K)$-classification problem and define

$$\textbf{label} = [0, 1, ..., N - 1] \tag{7}$$

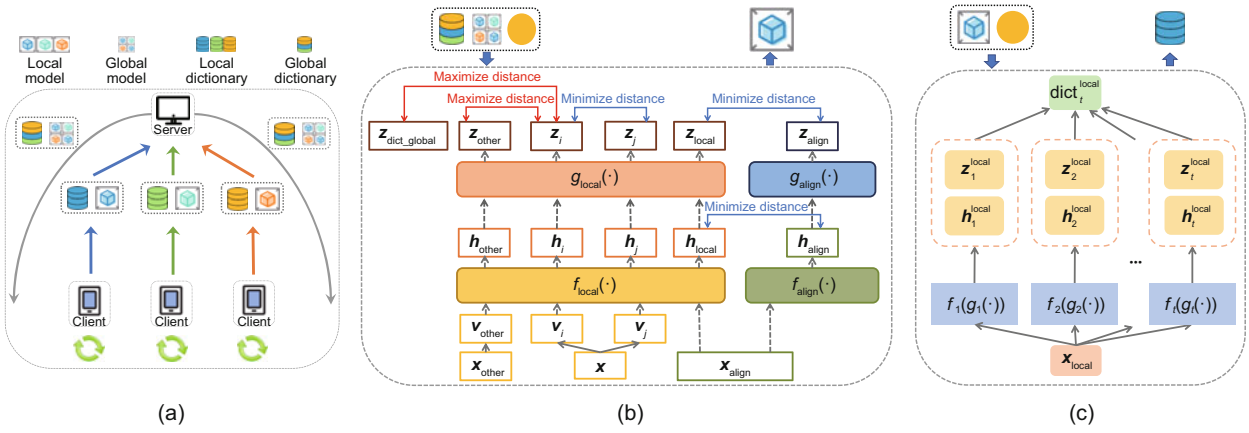

Fig. 3 Illustration of FedCA: (a) overview of FedCA (in each round, clients generate local models and dictionaries, and then the server gathers them to obtain the global model and dictionary); (b) local update of model (clients update local models by contrastive leaning with the dictionary and alignment modules); (c) local update of dictionary (clients generate local dictionaries via temporal ensembling). In (b), x_{other} is a sample different from sample x, x_{align} is a sample from the additional public dataset for alignment, f is the encoder, and g is the projection head

as a class indicator. Then the loss function is given as follows:

$$\text{loss}_{\text{contrastive}} = \text{CE}(\textbf{logits}/\tau, \textbf{label}), \tag{8}$$

where CE denotes the cross-entropy loss and τ is the temperature term.

Note that in each round, the shared dictionary is generated by the global model from the previous round, but the projections of local samples are encoded by current local models. The inconsistencies in representations may affect the function of the dictionary module, especially in a non-IID setting. We use temporal ensembling to alleviate this problem (Fig. 3c). To be specific, each client maintains a local ensemble dictionary consisting of projection set $\{\boldsymbol{Z}_{t-1}^i|\boldsymbol{x}_i \in D^u\}$. In each round, client u uses the trained local model to obtain projections $\{\boldsymbol{z}_t^i|\boldsymbol{x}_i \in D^u\}$ and accumulates them into ensemble dictionary by updating

$$\boldsymbol{Z}_t^i \leftarrow \alpha \boldsymbol{Z}_{t-1}^i + (1-\alpha)\boldsymbol{z}_t^i, \tag{9}$$

and then the normalized ensemble projection is given as

$$\tilde{\boldsymbol{z}}_t^i = \frac{\boldsymbol{Z}_t^i/(1-\alpha_t)}{||\boldsymbol{Z}_t^i/(1-\alpha_t)||_2} = \frac{\boldsymbol{Z}_t^i}{||\boldsymbol{Z}_t^i||_2}, \tag{10}$$

where $\alpha \in [0, 1)$ is a momentum parameter and $\boldsymbol{Z}_0^i = \boldsymbol{0}$.

We visualize the representations encoded by the local models trained via federated contrastive learning with the dictionary module in the same setting as the vanilla federated unsupervised approach. As shown in Fig. 2b, we find that the points of class 0 and class 1 are clustered in a small subspace of the representation space, which means that the dictionary module works well as we expected.

4.2 Alignment module for misalignment challenge

Due to the randomness in the training process, there might be differences between the representations generated by the two models trained on the same dataset, although these two models encode consistent spaces. The misalignment of representations may have an adverse effect on model aggregation.

To verify this, we use the angle between two representation vectors of the same image encoded by different models to measure the degree of difference in representations. Then we record the angles between representations generated by different local models in FL on CIFAR-10. We split the training data of CIFAR-10 into five IID sets randomly, and each set consists of 10 000 samples from all 10 classes. We randomly select two local models trained by the vanilla federated unsupervised approach (FedSimCLR is used as an example) and use them to obtain normalized representations on the test set of CIFAR-10. As shown in Fig. 4a, there is always a large difference in the angle (beyond 20°) between the representations encoded by the local models in the learning process.

We introduce an alignment module to tackle this challenge. As shown in Fig. 3b, we prepare an additional public dataset with a small size and train a model $g_{\text{align}}(f_{\text{align}}())$ (called the alignment model) on it. The local models are then trained via contrastive loss with a regularization term that replicates outputs of the alignment model on an alignment dataset. For a given client u, the loss functions are defined as follows:

$$\text{loss}^{h}_{\text{align}} = \sum_{i=1}^{|D_{\text{align}}|} ||\boldsymbol{h}_i^{\text{align}} - \boldsymbol{h}_i^{u}||_2^2, \tag{11}$$

$$\text{loss}^{z}_{\text{align}} = \sum_{i=1}^{|D_{\text{align}}|} ||\boldsymbol{z}_i^{\text{align}} - \boldsymbol{z}_i^{u}||_2^2, \tag{12}$$

$$\text{loss}_{\text{align}} = \text{loss}^{h}_{\text{align}} + \text{loss}^{z}_{\text{align}}, \tag{13}$$

where $\boldsymbol{h}_i^{\text{align}} = f_{\text{align}}(\boldsymbol{x}_i)$, $\boldsymbol{z}_i^{\text{align}} = g_{\text{align}}(\boldsymbol{h}_i^{\text{align}})$, $\boldsymbol{h}_i^{u} = f_u(\boldsymbol{x}_i)$, $\boldsymbol{z}_i^{u} = g_u(\boldsymbol{h}_i^{u})$, and $\boldsymbol{x}_i \in D_{\text{align}}$.

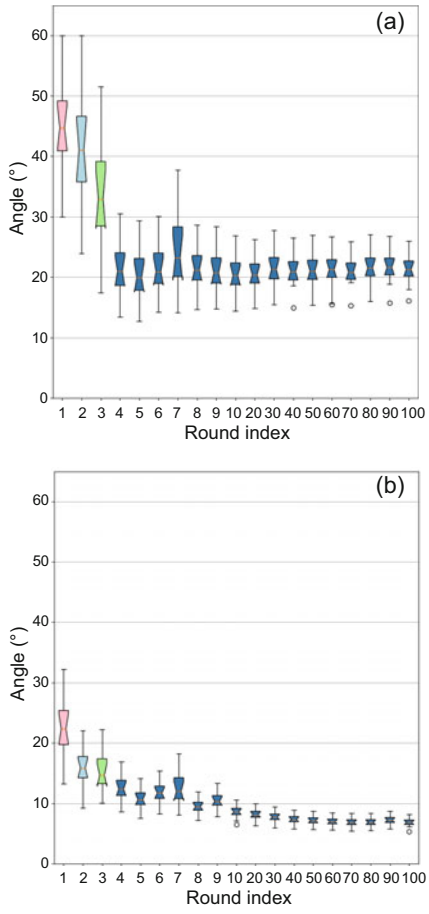

Fig. 4 Box plots of the angles between the representations encoded by local models on the CIFAR-10 dataset in FL with an IID setting: (a) FedSimCLR; (b) FedCA. FL: federated learning; IID: independent and identically distributed

We also calculate the angles between the representations of the local models trained via federated contrastive learning with the alignment module (3200 images sampled from the STL-10 dataset randomly are used for alignment) in the same setting as the vanilla federated unsupervised approach. As shown in Fig. 4b, the angles can be controlled within 10° after 10 training rounds, suggesting that the alignment module can help align the local models.

4.3 FedCA algorithm

From the above, the total loss function of the local model update is given as follows:

$$\text{loss} = \text{loss}_{\text{contrastive}} + \beta \text{loss}_{\text{align}}, \tag{14}$$

where β is a scale factor controlling the influence of the alignment module. Now we have a complete algorithm named FedCA, which can handle the challenges of FURL well, as shown in Fig. 3.

Algorithm 1 summarizes the proposed approach. In each round, clients update the local models with the contrastive loss and the alignment loss, and then generate local dictionaries. The server aggregates the local models into a global model and updates the global dictionary.

5 Experiments

FURL aims to learn a representation model from decentralized and unlabeled data. In this section, we present an empirical study of FedCA.

5.1 Experimental setup

5.1.1 Baselines

AutoEncoder is a generative method to learn representations in an unsupervised manner by generating a representation from the reduced encoding as close as possible to its original input (Hinton and Salakhutdinov, 2006). Predicting rotation is one of the proxy tasks of self-supervised learning by rotating samples by random multiples of 90° and predicting the degrees of rotations (Gidaris et al., 2018). We solely combine FedAvg with AutoEncoder (named FedAE), predicting rotation (named FedPR), and SimCLR (named FedSimCLR), separately, and use them as baselines for FURL.

Algorithm 1 Federated contrastive averaging with dictionary and alignment (FedCA)

Input: Client index u ($u = 1, 2, \ldots, n$), parameters of the global model (encoder and projection head) θ_t, parameters of the local model θ_t^u, global dictionary dict_t, local dictionary dict_t^u, proportion of selected clients C, number of local epochs E, local dataset D^u, and learning rate η

Server execution

Initialize θ_0
Prepare a public dataset D_{align} and an alignment model with parameters θ_{align}
for each round $t = 0, 1, ...,$ **do**
  $m \leftarrow \max(C \cdot n, 1)$
  $U_t \leftarrow$ random set of m clients
  for each client $u \in U_t$ **in parallel do**
    $\theta_{t+1}^u, \text{dict}_{t+1}^u \leftarrow \text{ClientUpdate}(u, \theta_t, \text{dict}_t)$
  end for
  $\theta_{t+1} \leftarrow \sum_{u \in U_t} \frac{|D^u|}{\sum_{i \in U_t} |D^i|} \theta_{t+1}^u$
  $\text{dict}_{t+1} \leftarrow \text{concat}\big([\{\text{dict}_{t+1}^u | u \in U_t\}], \text{dim} = 1\big)$
end for

ClientUpdate(u, θ, dict) // Run on client u

for each local epoch i from 1 to E **do**
  for batch $b \in D^u$ **do**
    $\theta^u \leftarrow \theta - \eta \bigtriangledown \mathcal{L}(\theta; b, \text{dict}, D_{\text{align}}, \theta_{\text{align}})$
    // Update θ with Eq. (14)
  end for
end for
Generate dict^u by Eqs. (9) and (10)
Return θ^u and dict^u

5.1.2 Datasets

The CIFAR-10/CIFAR-100 dataset (Krizhevsky, 2009) consists of 60 000 32×32 color images in 10/100 classes, with 6000/600 images per class, and there are 50 000 training images and 10 000 test images in CIFAR-10 and CIFAR-100. The MiniImageNet dataset (Deng et al., 2009; Vinyals et al., 2016) is extracted from the ImageNet dataset and consists of 60 000 84×84 color images in 100 classes. We split it into a training dataset with 50 000 samples and a test dataset with 10 000 samples. We implement FedCA and the baseline methods on the three datasets above in PyTorch (Paszke et al., 2019).

5.1.3 Federated setting

We deploy our experiments under a simulated FL environment, where we set a centralized node as the server and five distributed nodes as the clients. The number of local epochs (E) is five, and in each round, all of the clients obtain the global model and execute local training, i.e., the proportion of the selected clients $C = 1$. For each dataset, we consider two federated settings: IID and non-IID. Each client randomly samples 10 000 images from the entire training dataset in an IID setting, while in the non-IID setting, samples are split to clients by class, which means that each client has 10 000 samples of 2/20/20 classes of CIFAR-10/CIFAR-100/MiniImageNet.

5.1.4 Training details

We compare our approach with baseline methods on different encoders, including five-layer convolutional neural network (CNN) (Krizhevsky et al., 2012) and ResNet-50 (He KM et al., 2016). The encoder maps input samples to representations with 2048 dimensions, and then a multilayer perceptron (MLP) translates the representations to a vector with 128 dimensions used to calculate the contrastive loss. Adam is used as the optimizer, and the initial learning rate is 1×10^{-3} with 1×10^{-6} weight decay. We train models for 100 epochs with a minibatch size of 128. We set the dictionary size $K = 1024$, the momentum term of temporal ensembling $\alpha = 0.5$, and the scale factor $\beta = 0.01$. Furthermore, 3200 images randomly sampled from the STL-10 dataset are used for the alignment module. Data augmentation for contrastive representation learning includes random cropping and resizing, random color distortion, random flipping, and Gaussian blurring.

5.2 Evaluation protocols and results

5.2.1 Linear evaluation

We first study our method by linear classification on a fixed encoder to verify the representations learned in FURL. We perform FedCA and baseline methods to learn representations on CIFAR-10, CIFAR-100, and MiniImageNet without labels separately in a federated setting. Then, we fix the encoder and train a linear classifier with supervision on the entire dataset. We train this classifier with Adam as the optimizer for 100 epochs and report the top-1 classification accuracy on the test datasets of CIFAR-10, CIFAR-100, and MiniImageNet.

As shown in Table 2, federated averaging with contrastive learning works better than other

Table 2 Top-1 accuracies of algorithms for FURL on linear evaluation

Setting	Method	Accuracy (%)					
		CIFAR-10		CIFAR-100		MiniImageNet	
		Five-layer CNN	ResNet-50	Five-layer CNN	ResNet-50	Five-layer CNN	ResNet-50
IID	FedAE	61.23	65.47	34.07	36.56	28.21	31.97
	FedPR	55.75	63.52	29.74	30.89	24.76	26.63
	FedSimCLR	61.62	68.10	34.18	39.75	29.84	32.18
	FedCA (ours)	**64.87**	**71.25**	**39.47**	**43.30**	**35.27**	**37.12**
Non-IID	FedAE	60.14	63.74	33.94	37.27	29.00	30.44
	FedPR	54.94	60.31	30.70	32.39	24.74	25.91
	FedSimCLR	59.21	64.06	33.63	38.70	29.24	30.47
	FedCA (ours)	**63.02**	**68.01**	**38.94**	**42.34**	**34.95**	**35.01**

Values in bold are the best performance. FedAvg is combined with AutoEncoder (named FedAE), predicting rotation (named FedPR), and SimCLR (named FedSimCLR). CNN: convolutional neural network; FURL: federated unsupervised representation learning; IID: independent and identically distributed

unsupervised approaches. Moreover, our method outperforms all of the baseline methods due to the modules designed for FURL as we expected.

5.2.2 Semi-supervised learning

In federated scenarios, the private data at the clients may be only partly labeled, so we can learn a representation model without supervision and fine-tune it on labeled data. We assume that the ratios of labeled data of each client are 1% and 10%, separately. First, we train a representation model in FURL setting. Then, we fine-tune it (followed by an MLP consisting of a hidden layer and a rectified linear unit (ReLU) activation function) on labeled data for 100 epochs with Adam as the optimizer and a learning rate of 1×10^{-3}.

Table 3 reports the top-1 accuracy of various methods on CIFAR-10, CIFAR-100, and MiniImageNet. We observe that the accuracy of the global model trained by federated supervised learning on limited labeled data is significantly bad, and the use of the representation model trained in FURL as the initial model can improve performance relatively. Our method outperforms other approaches, suggesting that FURL benefits from the designed modules of FedCA, especially in a non-IID setting.

5.2.3 Transfer learning

A main goal of FURL is to learn a representation model from decentralized and unlabeled data for personalized downstream tasks. To verify whether the features learned in FURL are transferable, we set the models trained in FURL as the initial models, and then an MLP is used to be trained along with the encoder on other datasets. The image size of CIFAR ($32\times32\times3$) is resized to be the same as that in MiniImageNet ($84\times84\times3$) when we fine-tune the model learned from MiniImageNet on CIFAR. We train it for 100 epochs with Adam as the optimizer and set the learning rate to be 1×10^{-3}.

Table 4 shows that the model trained by FedCA achieves an excellent performance and outperforms all of the baseline methods in the non-IID setting.

5.3 Ablation study

5.3.1 Alignment and dictionary modules

We perform the ablation study analysis on CIFAR-10 in IID and non-IID settings to demonstrate the effectiveness of the alignment and dictionary modules (with temporal ensembling). We implement (1) FedSimCLR, (2) federated contrastive learning with only alignment module, (3) federated contrastive learning with only dictionary module, (4) federated contrastive learning with only dictionary module based on temporal ensembling, and (5) FedCA, and then a linear classifier is used to evaluate the performance of the frozen representation model with supervision. Fig. 5 shows the results.

We observe that the alignment module improves the performance by 1.4% in both IID and non-IID settings. With the help of the dictionary module (without temporal ensembling), there are 2.5% and 2.7% increases in the accuracy under the IID and non-IID settings, respectively. Moreover, we note

Table 3 Top-1 accuracies of algorithms for FURL on semi-supervised learning

Ratio of labeled data	Setting	Method	Accuracy (%)					
			CIFAR-10		CIFAR-100		MiniImageNet	
			Five-layer CNN	ResNet-50	Five-layer CNN	ResNet-50	Five-layer CNN	ResNet-50
1%	IID	FedAvg	31.84	26.68	9.35	8.09	5.83	5.42
		FedAE	35.98	36.86	13.36	14.53	11.71	12.84
		FedPR	34.51	36.47	13.15	14.20	11.52	12.34
		FedSimCLR	43.95	50.00	22.16	23.01	19.14	19.67
		FedCA (ours)	**45.05**	**50.67**	**22.37**	**23.32**	**19.20**	**20.22**
	Non-IID	FedAvg	20.99	17.72	6.22	5.37	3.92	3.03
		FedAE	23.08	23.43	9.96	9.63	8.45	8.43
		FedPR	22.83	23.17	9.83	9.38	8.30	8.58
		FedSimCLR	26.08	26.03	14.30	14.02	11.02	10.89
		FedCA (ours)	**28.96**	**28.50**	**17.02**	**16.48**	**13.39**	**13.03**
10%	IID	FedAvg	50.87	40.44	16.18	14.47	13.46	12.76
		FedAE	51.88	53.64	21.77	22.45	21.73	21.96
		FedPR	51.38	53.32	21.30	21.21	21.67	21.58
		FedSimCLR	59.27	60.67	31.11	31.56	28.45	28.79
		FedCA (ours)	**59.91**	**61.02**	**31.37**	**32.09**	**28.93**	**29.44**
	Non-IID	FedAvg	30.62	21.69	14.90	13.98	11.88	10.13
		FedAE	32.07	32.19	18.77	18.98	13.48	13.65
		FedPR	31.04	31.78	18.39	18.34	13.30	13.24
		FedSimCLR	32.52	33.83	19.91	20.01	15.90	16.03
		FedCA (ours)	**35.78**	**36.28**	**21.98**	**22.46**	**18.67**	**18.89**

Values in bold are the best performance. FedAvg is combined with AutoEncoder (named FedAE), predicting rotation (named FedPR), and SimCLR (named FedSimCLR). CNN: convolutional neural network; FURL: federated unsupervised representation learning; IID: independent and identically distributed

Table 4 Top-1 accuracies of algorithms for FURL on transfer learning

Setting	Method	Accuracy (%)					
		CIFAR-100 → CIFAR-10		MiniImageNet→CIFAR-10		MiniImageNet→CIFAR-100	
		Five-layer CNN	ResNet-50	Five-layer CNN	ResNet-50	Five-layer CNN	ResNet-50
	Random init	86.70	93.79	86.60	93.05	58.05	70.52
IID	FedAE	87.33	94.23	86.74	94.23	58.82	71.36
	FedPR	87.22	93.89	87.33	93.55	58.23	70.78
	FedSimCLR	87.80	94.88	**88.03**	94.87	**59.08**	71.85
	FedCA (ours)	**88.04**	**95.03**	87.91	**94.94**	58.91	**71.98**
Non-IID	FedAE	87.37	94.35	87.00	94.06	58.56	71.17
	FedPR	86.97	93.91	86.92	93.55	58.39	70.25
	FedSimCLR	87.04	94.02	86.81	93.97	58.11	70.91
	FedCA (ours)	**87.75**	**94.69**	**87.66**	**94.16**	**58.93**	**71.32**

Values in bold are the best performance. Random init means using the model with random initialization instead of pre-trained models. FedAvg is combined with AutoEncoder (named FedAE), predicting rotation (named FedPR), and SimCLR (named FedSimCLR). CNN: convolutional neural network; FURL: federated unsupervised representation learning; IID: independent and identically distributed

that the representation model learned in FURL benefits more from the temporal ensembling technique in the non-IID setting than in the IID setting, probably because the features learned in the IID setting are stable enough so that temporal ensembling plays a far less important role in the IID setting than in the non-IID setting. Fortunately, the model achieves excellent performance when we combine federated contrastive learning with the alignment and dictionary modules based on temporal ensembling, which

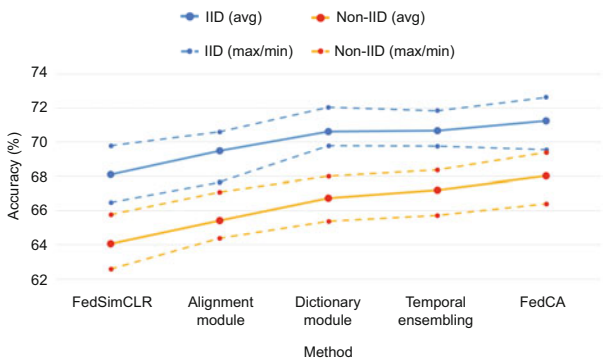

Fig. 5 Ablation study of modules designed for FURL by linear classification on CIFAR-10 (ResNet-50). FURL: federated unsupervised representation learning; IID: independent and identically distributed

suggests that both of these two modules can work collaboratively and help tackle the challenges in FURL.

5.3.2 Coefficient of alignment loss

To explore the effectiveness of the coefficient of alignment loss β, we run our algorithm on the CIFAR-10 dataset (IID setting, five-layer CNN) with different values of the hyper-parameter β.

The results are shown in Table 5. We can find that the values of β have a slight effect on the performance of the federated representation model. The reason for the performance differences may be that a small value of β cannot make the local models become aligned, so that the performance of the aggregated model will be degraded. A large value of β limits the function of the contrastive loss, so that the model ability cannot be guaranteed. We suggest that, in practice, people should select an appropriate value for β on a subset of data with a small size before the formal federated training.

Table 5 Ablation study for coefficient of alignment loss β

β	0.005	0.01	0.05	0.1
Top-1 accuracy (%)	64.16	64.87	64.34	63.93

6 Conclusions

We formulate a significant and challenging problem, termed federated unsupervised representation learning (FURL), and show the two main challenges (inconsistency of representation spaces and misalignment of representations). In this paper, we propose a contrastive learning based FL algorithm named FedCA, composed of the dictionary module and alignment module, to tackle the above challenges. Owing to these two modules, FedCA enables distributed local models to learn consistent and aligned representations while protecting data privacy. Our experimental results demonstrate that FedCA outperforms those algorithms that solely combine FL with unsupervised approaches and provides a stronger baseline for FURL.

In future work, we plan to extend FedCA to cross-modal scenarios where different clients may have data in different modes such as images, videos, texts, and audios.

Contributors

All authors contributed to the study conception and design. Fengda ZHANG, Chao WU, and Yueting ZHUANG proposed the motivation. Fengda ZHANG, Kun KUANG, and Long CHEN designed the method. Fengda ZHANG, Zhaoyang YOU, and Tao SHEN performed the experiments. Fengda ZHANG drafted the paper, and all authors commented on previous versions of the paper. Jun XIAO, Yin ZHANG, Fei WU, and Xiaolin LI revised the paper. All authors read and approved the final version.

Compliance with ethics guidelines

Fei WU and Yueting ZHUANG are editorial board members of *Frontiers of Information Technology & Electronic Engineering*. Fengda ZHANG, Kun KUANG, Long CHEN, Zhaoyang YOU, Tao SHEN, Jun XIAO, Yin ZHANG, Chao WU, Fei WU, Yueting ZHUANG, and Xiaolin LI declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in public repositories.

References

Baevski A, Zhou H, Mohamed A, et al., 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. Proc 34th Conf on Neural Information Processing Systems.

Bonawitz K, Ivanov V, Kreuter B, et al., 2017. Practical secure aggregation for privacy-preserving machine learning. Proc ACM SIGSAC Conf on Computer and Communications Security, p.1175-1191. https://doi.org/10.1145/3133956.3133982

Chen T, Kornblith S, Norouzi M, et al., 2020. A simple framework for contrastive learning of visual representations. Proc 37th Int Conf on Machine Learning, Article 149.

Chen XL, Fan HQ, Girshick R, et al., 2020. Improved baselines with momentum contrastive learning. https://arxiv.org/abs/2003.04297

Coates A, Ng AY, Lee H, 2011. An analysis of single-layer networks in unsupervised feature learning. Proc 14th Int Conf on Artificial Intelligence and Statistics, p.215-223.

Deng J, Dong W, Socher R, et al., 2009. ImageNet: a large-scale hierarchical image database. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.248-255. https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy A, Springenberg JT, Riedmiller M, et al., 2014. Discriminative unsupervised feature learning with convolutional neural networks. Proc 27th Int Conf on Neural Information Processing Systems, p.766-774.

Duan XY, Tang SL, Zhang SY, et al., 2018. Temporality-enhanced knowledge memory network for factoid question answering. *Front Inform Technol Electron Eng*, 19(1):104-115. https://doi.org/10.1631/FITEE.1700788

Gidaris S, Singh P, Komodakis N, 2018. Unsupervised representation learning by predicting image rotations. Proc 6th Int Conf on Learning Representations.

Hadsell R, Chopra S, LeCun Y, 2006. Dimensionality reduction by learning an invariant mapping. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.1735-1742. https://doi.org/10.1109/CVPR.2006.100

Hassani K, Ahmadi AHK, 2020. Contrastive multi-view representation learning on graphs. Proc 37th Int Conf on Machine Learning, p.4116-4126.

He CY, Yang ZY, Mushtaq E, et al., 2021. SSFL: tackling label deficiency in federated learning via personalized self-supervision. https://arxiv.org/abs/2110.02470

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

He KM, Fan HQ, Wu YX, et al., 2020. Momentum contrast for unsupervised visual representation learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9729-9738. https://doi.org/10.1109/CVPR42600.2020.00975

Hinton GE, Salakhutdinov RR, 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504-507. https://doi.org/10.1126/science.1127647

Jeong E, Oh S, Kim H, et al., 2018. Communication-efficient on-device machine learning: federated distillation and augmentation under non-IID private data. https://arxiv.org/abs/1811.11479v1

Ji SX, Saravirta T, Pan SR, et al., 2021. Emerging trends in federated learning: from model fusion to federated X learning. https://arxiv.org/abs/2102.12920

Jin YL, Wei XG, Liu Y, et al., 2020. Towards utilizing unlabeled data in federated learning: a survey and prospective. https://arxiv.org/abs/2002.11545

Kairouz P, McMahan HB, Avent B, et al., 2021. Advances and open problems in federated learning. https://arxiv.org/abs/1912.04977

Kempe D, McSherry F, 2008. A decentralized algorithm for spectral analysis. *J Comput Syst Sci*, 74(1):70-83. https://doi.org/10.1016/j.jcss.2007.04.014

Kingma DP, Welling M, 2014. Auto-encoding variational Bayes. Proc 2nd Int Conf on Learning Representations.

Konečný J, McMahan HB, Yu FX, et al., 2017. Federated learning: strategies for improving communication efficiency. https://arxiv.org/abs/1610.05492

Krizhevsky A, 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009, University of Toronto, Toronto, Canada.

Krizhevsky A, Sutskever I, Hinton GE, 2012. ImageNet classification with deep convolutional neural networks. Proc 25th Int Conf on Neural Information Processing Systems, p.1097-1105.

Kuang K, Li L, Geng Z, et al., 2020. Causal inference. *Engineering*, 6(3):253-263. https://doi.org/10.1016/j.eng.2019.08.016

Lei N, An DS, Guo Y, et al., 2020. A geometric understanding of deep learning. *Engineering*, 6(3):361-374. https://doi.org/10.1016/j.eng.2019.09.010

Li QB, He BS, Song D, 2021. Model-contrastive federated learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10713-10722. https://doi.org/10.1109/CVPR46437.2021.01057

Li T, Sahu AK, Zaheer M, et al., 2020. Federated optimization in heterogeneous networks. Proc 3rd MLSys Conf.

Liang JL, Zhang MH, Zeng XY, et al., 2014. Distributed dictionary learning for sparse representation in sensor networks. *IEEE Trans Image Process*, 23(6):2528-2541. https://doi.org/10.1109/TIP.2014.2316373

Logeswaran L, Lee H, 2018. An efficient framework for learning sentence representations. Proc 6th Int Conf on Learning Representations.

Lyu YG, 2020. Artificial intelligence: enabling technology to empower society. *Engineering*, 6(3):205-206. https://doi.org/10.1016/j.eng.2020.01.005

McMahan B, Moore E, Ramage D, et al., 2017. Communication-efficient learning of deep networks from decentralized data. Proc 20th Int Conf on Artificial Intelligence and Statistics, p.1273-1282.

Mikolov T, Sutskever I, Chen K, et al., 2013. Distributed representations of words and phrases and their compositionality. Proc 26th Int Conf on Neural Information Processing Systems, p.3111-3119.

Pan YH, 2020. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Paszke A, Gross S, Massa F, et al., 2019. PyTorch: an imperative style, high-performance deep learning library. Proc 33rd Conf on Neural Information Processing Systems, p.8026-8037.

Pathak D, Agrawal P, Efros AA, et al., 2017. Curiosity-driven exploration by self-supervised prediction. Proc IEEE Conf on Computer Vision and Pattern Recognition Workshops, p.16-17. https://doi.org/10.1109/CVPRW.2017.70

Qiu JZ, Chen QB, Dong YX, et al., 2020. GCC: graph contrastive coding for graph neural network pre-training. Proc 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining, p.1150-1160. https://doi.org/10.1145/3394486.3403168

Radford A, Metz L, Chintala S, 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. Proc 4th Int Conf on Learning Representations.

Raja H, Bajwa WU, 2016. Cloud K-SVD: a collaborative dictionary learning algorithm for big, distributed data. *IEEE Trans Signal Process*, 64(1):173-188. https://doi.org/10.1109/TSP.2015.2472372

Sattler F, Wiedemann S, Müller KR, et al., 2020. Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Trans Neur Netw Learn Syst*, 31(9):3400-3413. https://doi.org/10.1109/TNNLS.2019.2944481

Sattler F, Korjakow T, Rischke R, et al., 2021. FEDAUX: leveraging unlabeled auxiliary data in federated learning. *IEEE Trans Neur Netw Learn Syst*, early access. https://doi.org/10.1109/TNNLS.2021.3129371

Sermanet P, Lynch C, Chebotar Y, et al., 2018. Time-contrastive networks: self-supervised learning from video. Proc IEEE Int Conf on Robotics and Automation, p.1134-1141. https://doi.org/10.1109/ICRA.2018.8462891

Shakeri Z, Raja H, Bajwa WU, 2014. Dictionary learning based nonlinear classifier training from distributed data. Proc IEEE Global Conf on Signal and Information Processing, p.759-763. https://doi.org/10.1109/GlobalSIP.2014.7032221

Shi HZ, Zhang YC, Shen ZJ, et al., 2022. Federated self-supervised contrastive learning via ensemble similarity distillation. https://arxiv.org/abs/2109.14611v1

Sohn K, 2016. Improved deep metric learning with multi-class N-pair loss objective. Proc 30th Int Conf on Neural Information Processing Systems, p.1857-1865.

Tian YL, Krishnan D, Isola P, 2020. Contrastive multiview coding. Proc 16th European Conf on Computer Vision, p.776-794. https://doi.org/10.1007/978-3-030-58621-8_45

van Berlo B, Saeed A, Ozcelebi T, 2020. Towards federated unsupervised representation learning. Proc 3rd ACM Int Workshop on Edge Systems, Analytics and Networking, p.31-36. https://doi.org/10.1145/3378679.3394530

van den Oord A, Li YZ, Vinyals O, 2019. Representation learning with contrastive predictive coding. https://arxiv.org/abs/1807.03748

Vinyals O, Blundell C, Lillicrap T, et al., 2016. Matching networks for one shot learning. Proc 30th Int Conf on Neural Information Processing Systems, p.3637-3645.

Wang HY, Yurochkin M, Sun YK, et al., 2020. Federated learning with matched averaging. Proc 8th Int Conf on Learning Representations.

Wang TZ, Isola P, 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. Proc 37th Int Conf on Machine Learning, p.9929-9939.

Wu SX, Wai HT, Li L, et al., 2018. A review of distributed algorithms for principal component analysis. *Proc IEEE*, 106(8):1321-1340. https://doi.org/10.1109/JPROC.2018.2846568

Wu YW, Zeng DW, Wang ZP, et al., 2021. Federated contrastive learning for volumetric medical image segmentation. Proc 24th Int Conf on Medical Image Computing and Computer-Assisted Intervention, p.367-377. https://doi.org/10.1007/978-3-030-87199-4_35

Wu ZR, Xiong YJ, Yu SX, et al., 2018. Unsupervised feature learning via non-parametric instance discrimination. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3733-3742. https://doi.org/10.1109/CVPR.2018.00393

Yang Q, Liu Y, Chen TJ, et al., 2019. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol*, 10(2):12. https://doi.org/10.1145/3298981

Yang ZL, Dai ZH, Yang YM, et al., 2019. XLNet: generalized autoregressive pretraining for language understanding. Proc 33rd Int Conf on Neural Information Processing Systems, Article 517.

Zhao Y, Li M, Lai LZ, et al., 2022. Federated learning with non-IID data. https://arxiv.org/abs/1806.00582

Zhou LK, Tang SL, Xiao J, et al., 2017. Disambiguating named entities with deep supervised learning via crowd labels. *Front Inform Technol Electron Eng*, 18(1):97-106. https://doi.org/10.1631/FITEE.1601835

Zhu YX, Gao T, Fan LF, et al., 2020. Dark, beyond deep: a paradigm shift to cognitive AI with humanlike common sense. *Engineering*, 6(3):310-345. https://doi.org/10.1016/j.eng.2020.01.011

Zhuang WM, Gan X, Wen YG, et al., 2021a. Collaborative unsupervised visual representation learning from decentralized data. Proc IEEE/CVF Int Conf on Computer Vision, p.4892-4901. https://doi.org/10.1109/ICCV48922.2021.00487

Zhuang WM, Wen YG, Zhang S, 2021b. Joint optimization in edge-cloud continuum for federated unsupervised person re-identification. Proc 29th ACM Int Conf on Multimedia, p.433-441. https://doi.org/10.1145/3474085.3475182

Zhuang WM, Wen YG, Zhang S, 2022. Divergence-aware federated self-supervised learning. Proc 10th Int Conf on Learning Representations.

Zhuang YT, Wu F, Chen C, et al., 2017. Challenges and opportunities: from big data to knowledge in AI 2.0. *Front Inform Technol Electron Eng*, 18(1):3-14. https://doi.org/10.1631/FITEE.1601883

Zhuang YT, Cai M, Li XL, et al., 2020. The next breakthroughs of artificial intelligence: the interdisciplinary nature of AI. *Engineering*, 6(3):245-247. https://doi.org/10.1016/j.eng.2020.01.009