Learn from Others and Be Yourself in Heterogeneous Federated Learning

Wenke Huang[1], Mang Ye[1,2*], Bo Du[1,2*]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, Wuhan, China,
[2] Hubei Luojia Laboratory, Wuhan, China,

https://github.com/WenkeHuang/FCCL

Abstract

Federated learning has emerged as an important distributed learning paradigm, which normally involves collaborative updating with others and local updating on private data. However, heterogeneity problem and catastrophic forgetting bring distinctive challenges. First, due to non-i.i.d (identically and independently distributed) data and heterogeneous architectures, models suffer performance degradation on other domains and communication barrier with participants models. Second, in local updating, model is separately optimized on private data, which is prone to overfit current data distribution and forgets previously acquired knowledge, resulting in catastrophic forgetting. In this work, we propose FCCL (Federated Cross-Correlation and Continual Learning). For heterogeneity problem, FCCL leverages unlabeled public data for communication and construct cross-correlation matrix to learn a generalizable representation under domain shift. Meanwhile, for catastrophic forgetting, FCCL utilizes knowledge distillation in local updating, providing inter and intra domain information without leaking privacy. Empirical results on various image classification tasks demonstrate the effectiveness of our method and the efficiency of modules.

1. Introduction

Deep learning algorithms have achieved remarkable progress, owing to the availability of large-scale data [8, 51, 69]. However, in the real world, data are commonly dispersed over different participants (*e.g.*, mobile devices, organizations). Due to growing privacy concerns and strict data protection regulations [84], participants cannot integrate data together to train a model. Driven by such realistic issues, federated learning [33, 34, 58, 59, 89] provides a privacy-preserving paradigm, where participants collaboratively learn a model without leaking private data. It has been an active and challenging research topic and shows promising results in real-world setting [17, 19, 29, 52, 54].

*Corresponding Author: Mang Ye, Bo Du

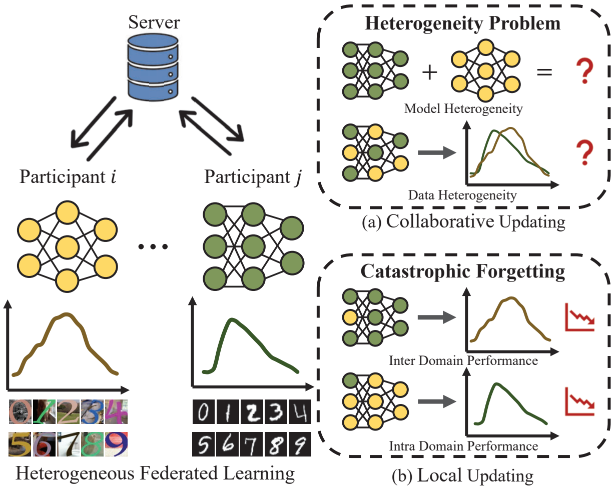


Figure 1. **Problem illustration of heterogeneous federated learning.** (a) In collaborative updating, how to handle communication problem of heterogeneous models and learn a generalizable representation under heterogeneous data (domain shift)? (b) In local updating, how to alleviate catastrophic forgetting to present stable and satisfactory performance in both inter and intra domains?

Along with its pilot progress, researches on federated learning are baffled by some key challenges [30, 42]. An inevitable and practical challenge is heterogeneity problem. On the one hand, distributed data might be non-i.i.d (identically and independently distributed), leading to **data heterogeneity** [30,39,95]. A myriad of methods [43,46,73,77] incorporate extra proximal terms to handle the data in *label distribution skew* (prior probability shift) [30], neglecting the fact that there exists *domain shift* (same label, different features) [60, 64, 66]. In particular, private model suffers severe performance degradation on other domains with noticeably different distribution. As a result, learning a generalizable representation under domain shift is technically challenging. On the other hand, due to different design criteria, distinct hardware capabilities [20, 86] and intellectual property rights [56], participants require to customize models, which poses a practical challenge: **model heterogeneity**. Preceding methods are developed under the assumption that local models share parameters or gradients, which cannot work on heterogeneous models. In order to solve this problem, a main stream of subsequent effort leverages knowledge transfer through labeled data [38, 74], shared model [48, 72, 92] or group operation [21, 50]. But these methods have different limitations. Specifically, labeled data require server to collect data with similar distributions to private data, which causes costly human efforts and needs special domain expertise. For shared model, it raises computational cost and necessitates additional model structure in participant side. Group operation leverages unlabeled public data to measure distribution divergence. However, these methods mainly focus on label distribution skew and consider the performance on one domain. Simultaneously considering data and model heterogeneity, an essential issue has long been overlooked: (a) *How to learn a generalizable representation in heterogeneous federated learning?*

Besides heterogeneity problem, another impediment for federated learning steams from its paradigm. Generally, federated learning could be viewed as a two-step cyclic process: *collaborative updating* and *local updating* [58, 89]. In *collaborative updating*, participants learn from others. In *local updating*, model is optimized on private data, which is prone to overfit current knowledge and forget previous knowledge, resulting in **catastrophic forgetting** [57]. To tackle this challenge, one type of methods typically performs fine-tuning for several rounds [38, 50, 58, 74, 88]. However, carefully configuring hyper-parameters to achieve satisfactory performance is time-consuming and cannot tackle this problem systematically. Current popular solutions [41,43,73,77] focus on calculating parameter stiffness to regulate models, which can not explicitly depict the degree of effect from different participants. Consequently, a natural question arises: (b) *How to balance multiple knowledge to reduce catastrophic forgetting?* We further explain heterogeneity problem and catastrophic forgetting in Fig. 1.

For the heterogeneity problem, we take inspiration from the self-supervised learning [5, 6, 11, 13, 18, 25, 49, 91, 94]. In particular, self-supervised learning aims to learn a generalizable representation through rich and diverse data for downstream tasks and unseen classes. Intuitively, we expect that the models would present similar logits output for the same classes in different domains. This motivates us to leverage unlabeled public data for *Federated Cross-Correlation Learning*, which is diverse and easy to obtain. Specifically, we try to maximize the similarity between its output and minimize the redundancy within logits output on unlabeled public data. Through correlating same dimensions and decorrelating different dimensions on logits output, models would learn class invariance and encourage the diversity of different classes. Thus, our method handles the communication problem in heterogeneous models and learns a generalizable representation under domain shift.

To handle catastrophic forgetting, we develop *Federated Continual Learning* via knowledge distillation [2, 24] in local updating to continually learn from inter and intra domains. To avoid forgetting inter domain information in local updating stage, we propose to distill the knowledge of intra-domain (local) model learned in previous rounds, where it captured the inter domain information after communication with other participants. In addition, for intra domain forgetting problem, we leverage the initially pretrained local model (without knowledge learned from others) to constrain the later local updating for each participant. Therefore, balancing knowledge through distillation with these two models is reasonable to handle the catastrophic forgetting.

In this work, we propose a novel federated learning method, dubbed *FCCL* (**F**ederated **C**ross-Correlation and **C**ontinual **L**earning). The overview of *FCCL* is illustrated in Fig. 2. In a nutshell, our contributions are three-fold:

- We formulate a simple and effective method for heterogeneous federated learning. Through leveraging unlabeled public data and adopting self-supervised learning, heterogeneous models achieve communication and learn a generalizable representation.
- We explore to alleviate catastrophic forgetting in federated learning. Through inter and intra domain knowledge distillation with updated and pretrained models, it balances knowledge from others and itself.
- We conduct extensive experiments on two image classification tasks (*e.g.*, *Digits* [27, 37, 62, 68] and *Office-Home* [82]) with unlabeled public data [35, 69, 87]. *FCCL* achieves superior inter and intra domain performance over related methods. Ablation study on core module validates its efficacy and indispensability.

2. Related Work

Federated with Data Heterogeneity. A pioneering work proposed the currently most widely used algorithm, *FedAVG* [58]. But it suffers performance deterioration on non-i.i.d data (data heterogeneity). Shortly after, a large body methods [12, 41, 43, 73, 77] research on non-i.i.d data. These methods mainly focus on label distribution skew, where non-i.i.d data [30] are formed by partitioning existing data based on label space with limited domain shift. However, when private data sampled from different data domains, these works do not consider inter domain performance but only focus on learning an internal model. Latest researches have studied related problems of unsupervised domain adaptation for target domain [45, 65] and domain generalization on unseen domains [52]. However, collecting data in the target domain can be time-consuming and impractical. Meanwhile, considering the performance on unknown domains is an idealistic setting. For more realistic settings, participants are probably more interested in the performance on other domains, which could directly improve economic benefits. In this work, we focus on improving inter domain performance under domain shift.

Federated with Model Heterogeneity. With the demand for unique models, federated learning with model heterogeneity has been an active area of research. *FedMD* [38], *CRONUS* [4] and *CFD* [71] operate on labeled public data (with similar distribution) via knowledge distillation [2, 24]. Therefore, these approaches heavily rely on the quality of labeled public data, which may not always be available on the server. Latest works (*e.g.*,*FedDF* [50], *FedKT* [40] and *FEDGEN* [75]) have proven the feasibility to do distillation on unlabeled public data or synthetic data. However, these methods leverage unlabeled public data to reach semantic information consistency through various measuring metrics [9, 36], which are not suitable to learn a generalizable representation and and thus lead to a bad inter domain performance. Another direction is introducing shared extra model such as *FML* [72] and *LG-FEDAVG* [48]. However, these techniques may not be applicable when considering the additional computing overhead and expensive communication cost. In this paper, based on unlabeled public data, we correlate same dimensions and decorrelate different dimensions to learn a generalizable representation in heterogeneous federated learning.

Self-Supervised Learning. Self-supervised Learning has emerged as a powerful method for learning useful representation without supervision from labels, largely reducing the performance gap between supervised models on various downstream vision tasks. Many related methods rely on contrastive learning (*e.g.*, *SimCLR* [5], *MoCo* [7, 22]), which contrast positive pairs against negative pairs and minimizes difference between positive pairs for avoiding collapsing solutions [79, 90]. Recently, another line of works (*e.g.*, *BYOL* [15], *SimSiam* [6]) employs asymmetry of the learning update (stop-gradient operation) to avoid trivial solutions. Besides, some methods (*e.g.*, *W-MSE* [3], *Barlow Twins* [91])) investigate the possibility of feature decorrelation based on *Cholesky Decomposition* [83] and *Information Bottleneck* [80]. There are several works that consider federated learning with self-supervised learning (*e.g.*, *FURL* [93], *MOON* [41]). They focus on the unsupervised learning setting and label distribution skew with model homogeneity respectively. The key difference between *FCCL* and above self-supervised learning methods is that ours is designed for federated setting rather centralized setting. Inspired by self-supervised learning, *FCCL* constructs the comparison between different models in federated learning.

Catastrophic Forgetting. Catastrophic forgetting has been an essential problem in continual learning when models continuously learn from a stream data, with the goal of gradually extending acquired knowledge and using it for future learning [14, 57]. The challenge lies in the continuously changing class distributions of each task [63, 81]. Existing continual learning works on tackling catastrophic forgetting can be broadly divided into three branches [10]: replay methods [1, 67], regularization-based methods [32, 47, 53, 85] and parameter isolation methods [55, 61, 70]. As for federated learning, data are distributed rather than sequential like continual learning. But these differences aside, both continual learning and federated learning share a common challenge - how to balance the knowledge from different data distribution. Unlike continual learning methods, we focus on alleviate catastrophic forgetting in distributed data rather than time series data. In particular, we expect to balance and boost both inter and intra domain performance.

3. Method

Problem Setup and Notations. Following the standard federated learning setup, there are K participants (indexed by i). Each participant has a local model θ_i and private data $D_i = \{(X_i, Y_i)|X_i \in \mathbb{R}^{N_i \times D}, Y_i \in \mathbb{R}^{N_i \times C}\}$, where N_i denotes the number of private data, D represents input size and C is defined as the number of classes for classification. Meanwhile, the private data distribution is denoted as $P_i(X, Y)$ and rewritten as $P_i(X|Y)P_i(Y)$. Furthermore, in heterogeneous federated learning, **data heterogeneity** and **model heterogeneity** are defined as following:

- *Data heterogeneity*: $P_i(X|Y) \neq P_j(X|Y)$. There exists domain shift among private data, *i.e.*, conditional distribution $P(X|Y)$ of private data vary across participants even if $P(Y)$ is shared. Specifically, same label Y has distinctive feature X in different domains.

- *Model heterogeneity*: $Shape(\theta_i) \neq Shape(\theta_j)$. Participants customize models independently, *i.e.*, for classification task, the selected backbones (*e.g.*, *ResNet* [23], *EfficientNet* [78] and *MobileNet* [26]) are different with differential classifier models.

We leverage unlabeled public data $D_0 = \{X_0|X_0 \in \mathbb{R}^{N_0 \times D}\}$ to realize communication. The public data are relatively easy to access in real scenarios, *e.g.*, existing datasets [8, 51, 69] and web images [44]. The goal for i^{th} participant is to reach communication and learn a model θ_i with generalizable representation. In addition, considering

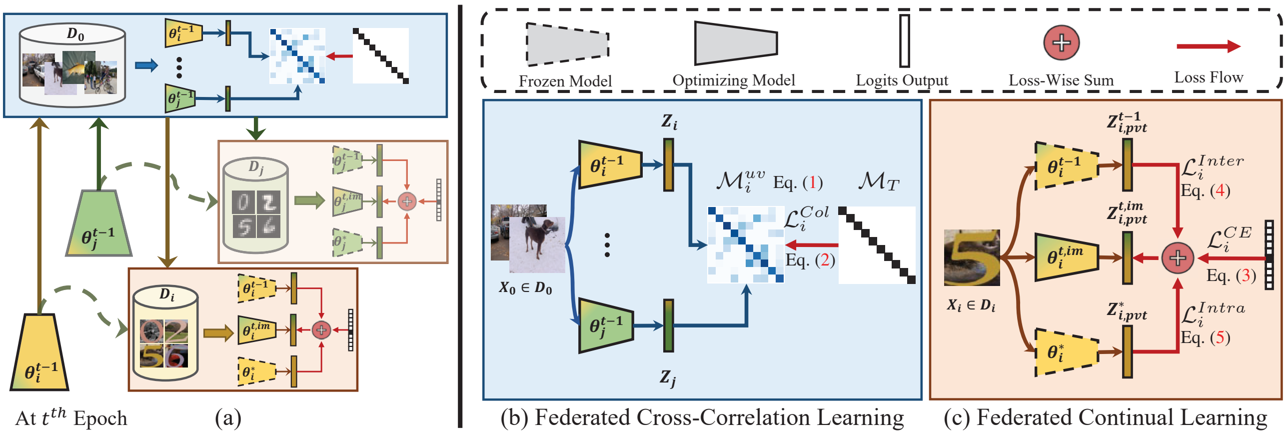


Figure 2. **Illustration of *FCCL*.** (a) Simplified schematization of our method that solves heterogeneity problem and catastrophic forgetting via *Federated Cross-Correlation Learning* and *Federated Continual Learning*. (b) *Federated Cross-Correlation Learning* § 3.1: Construct cross-correlation matrix $\mathcal{M}_i$ to target matrix, $\mathcal{M}_T = 2 \times eye(C) - ones(C)$, where on-diagonal is 1, off-diagonal is -1. (c) *Federated Continual Learning* § 3.2: Distillation with updated and pretrained models offers inter and intra domain knowledge without privacy leaking. The gradient color proportion reflects the degree of influence by other participants. Best viewed in color. Zoom in for details.

catastrophic problem, θ_k is required to present both higher and stabler inter and intra domain performance.

Overview of Framework. The framework of our method is illustrated in Fig. 2. Specifically, in collaborative updating, we measure cross-correlation matrix between logits output on unlabeled public data to make similarity and reduce redundancy. Meanwhile, in local updating, ours continually balances multi domains information through knowledge distillation. Next, we will first describe *Federated Cross-Correlation Learning* § 3.1. Then we introduce *Federated Continual Learning* § 3.2.

3.1. Federated Cross-Correlation Learning

Motivation of Dimension-Level Operation. Motivated by the success of self-supervised learning via *Information Bottleneck* [80, 91], a generalizable representation should be as informative as possible about image, while being as invariant as possible to the specific domain distortions that are applied to this sample. In our work, domain shift results in distinctive feature X for the same label Y in different domains. Therefore, the distribution of logits output along the batch dimension on different domains is not identical. Moreover, different dimensions of logits output are corresponding to distinct classes. Thus, we need to encourages the invariance of same dimensions and the diversity of different dimensions. Private data carries specific domain information and is under privacy protection, which is not suitable and feasible to do self-supervised learning. Therefore, we leverage the unlabeled public data, which are normally generated and collected from multi domains and is easy to obtain. We optimize private models through requiring logits output invariant to domain distortion and decorrelating different dimensions of logits output on unlabeled public data.

Construction of Cross-Correlation Matrix. Specifically, we get the logits output for i^{th} participant: $Z_i = f(\theta_i, X_0) \in \mathbb{R}^{N_0 \times C}$. For i^{th} and j^{th} participant, the logits output on unlabeled public data is Z_i and Z_j. Notably, considering the computing burden on the server side, we calculate average logits output: $\overline{Z} = \frac{1}{K}\sum_i Z_i$. Then, we compute cross-correlation matrix, $\mathcal{M}_i$ for i^{th} participant with average logits output as:

$$\mathcal{M}_i^{uv} \triangleq \frac{\sum_b ||Z_i^{b,u}||\ ||\overline{Z}^{b,v}||}{\sqrt{\sum_b ||Z_i^{b,u}||^2}\sqrt{\sum_b ||\overline{Z}^{b,v}||^2}}. \quad (1)$$

where b indexes batch samples, u, v index the dimension of logits output and $||\cdot||$ is the normalization operation along the batch dimension. $\mathcal{M}_i$ is a square matrix with size of output dimensionality, C and values comprised between -1 (*i.e.*, dissimilarity) and 1 (*i.e.*, similarity). Then, collaborative loss for i^{th} participant is defined as:

$$\mathcal{L}_i^{Col} \triangleq \sum_u (1 - \mathcal{M}_i^{uu})^2 + \lambda_{Col} \sum_u \sum_{v \neq u} (1 + \mathcal{M}_i^{uv})^2, \quad (2)$$

where λ_{Col} is a positive constant trading off the importance of the first and second terms of loss. Naturally, when on-diagonal terms of the cross-correlation matrix take the value $+1$, it encourages the logits output from different participants to be similar; when off-diagonal terms of the cross-correlation matrix take value -1, it encourages the diversity of logits output, since different dimensions of these logits output will be uncorrelated to each other.

Comparison with Analogous Methods. *FedMD* [38] relies on minimizing mean square error on annotated data. *FedDF* [50] reaches logits output distribution consistency on unlabeled public data. However, in our work, we expect

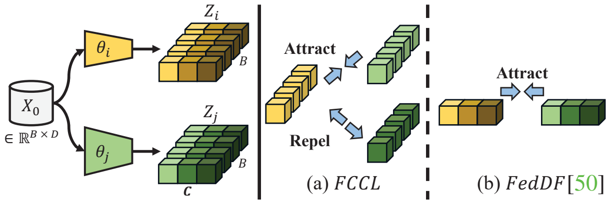


Figure 3. **Conceptual comparison.** The unlabeled public data X_0 with batch size B and input size D are fed into different models. The logits output has C dimensions. (a) *FCCL* learns invariance in same dimensions and decorrelates pairs of different dimensions on the batch-wise normalized logits output in Eq. (1). (b) *FedDF* [50] calculates the distribution divergence where instance-wise normalized logits output is compared inside a batch.

to achieve correlation of same dimensions but decorrelation of different dimensions on unlabeled public data. Besides, we do operation along the batch dimension, which means that we view unlabeled public data as ensemble rather than individual sample. It is advantageous to eliminate anomalous sample disturbance. We further illustrate the conceptual comparison between *FCCL* and *FedDF* in Fig. 3.

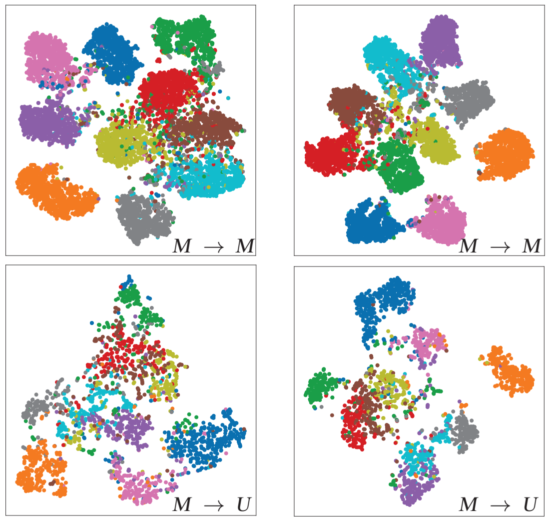
3.2. Federated Continual Learning

Typical Supervision Loss. For local updating in federated learning, current methods [38, 50, 58, 74] typically cast this process as a supervised classification problem. Specifically, at t^{th} communication round, after the collaborative updating, the i^{th} private model is defined as $\theta_i^{t,im}$. Then, optimize $\theta_i^{t,im}$ on private data $D_i(X_i, Y_i)$ for fixed epochs. Given the logits output $Z_{i,pvt}^{t,im} = f(\theta_i^{t,im}, X_i)$ for private data X_i w.r.t its ground truth label Y_i, the cross-entropy loss is optimized with $\texttt{softmax}$:

$$\mathcal{L}_i^{CE} = -\mathbf{1}_{Y_i} \log(\texttt{softmax}(Z_{i,pvt}^{t,im}), \tag{3}$$

where $\mathbf{1}_{Y_i}$ denotes the one-hot encoding of Y_i and $\texttt{softmax}(Z_{i,pvt}^{t,im}) = \frac{\exp(Z_{i,pvt}^{t,im})}{\sum_{c'=1}^{C} \exp(Z_{i,pvt}^{t,im,c'})}$. Such training objective design would suffer catastrophic forgetting mainly due to the following two limitations: **1)** In local updating, without supervision from other participants, models easily overfit current data distribution and present poor inter domain performance. **2)** Besides, it only penalizes the prediction independently with prior probabilities, which provides limited and hard intra domain information [24].

Dual-Domain Knowledge Distillation Loss. In this work, we develop a federated continual learning method to address both **1)** and **2)** through regularizing the objective from model-wise aspect. Specifically, at the end of $t-1^{th}$ round, the updated model, θ_i^{t-1} involves the knowledge learned from other participants. We calculate the logits output on private data: $Z_{i,pvt}^{t-1} = f(\theta_i^{t-1}, X_i)$. The inter domain knowledge distillation loss is defined as:



(a) CE (b) Ours

Figure 4. **Visualization of features** learned with (left) typical supervision loss (*i.e.*, cross-entropy loss, $\mathcal{L}^{CE}$ in Eq. (3) and (right) optimization objective based on dual-domain knowledge distillation (*i.e.*, $\mathcal{L}^{Dual}$ in Eq. (6)) on intra (top) and inter (bottom) domain. M and U represent *MNIST* and *USPS* respectively. Features are colored based on class labels.

$$\mathcal{L}_i^{Inter} = \sigma(Z_{i,pvt}^{t-1}) \log \frac{\sigma(Z_{i,pvt}^{t-1})}{\sigma(Z_{i,pvt}^{t,im})}, \tag{4}$$

where σ denote $\texttt{softmax}$ function. As Eq. (4), the purpose is to continually learn from others while preserving privacy, so as to guarantee inter domain performance and handle catastrophic forgetting in federated learning. Besides, for the i^{th} participant, it is feasible to pretrain a model, θ_i^* on private data. We measure the logits output on private data: $Z_{i,pvt}^* = f(\theta_i^*, X_i)$. The intra domain knowledge distillation loss can be given as :

$$\mathcal{L}_i^{Intra} = \sigma(Z_{i,pvt}^*) \log \frac{\sigma(Z_{i,pvt}^*)}{\sigma(Z_{i,pvt}^{t,im})}. \tag{5}$$

Knowledge distillation with pretrained model provides soft and rich intra domain information. Further, it cooperates with the former typical supervision loss (*i.e.*, cross-entropy loss) in Eq. (3) to provide soft and hard intra domain information to ensure intra domain performance. To some extent, above two models (*i.e.* updated model θ_i^{t-1} and pretrained model θ_i^*) respectively represent inter and intra 'teacher' model. Through knowledge distillation, balancing knowledge from others and itself simultaneously boosts both inter and intra domain performance. The dual-domain knowledge distillation is calculated by

$$\mathcal{L}_i^{Dual} = \mathcal{L}_i^{Inter} + \mathcal{L}_i^{Intra}. \tag{6}$$

Algorithm 1: The *FCCL* Framework

Input: Communication rounds T, local epochs E, participants number K, unlabeled public data (X_0), i^{th} private data (X_i, Y_i) and pretrained model θ_i^*, hyper-parameter λ_{Col}, λ_{Loc}

for $t = 1, 2, ..., T$ **do**
  for $i = 1, 2, ..., K$ **do**
    Calculate logits output: $Z_i = f(\theta_i^{t-1}, X_0)$
  Average logits output: $\overline{Z} = \frac{1}{K}\sum_i Z_i$
  for $i = 1, 2, ..., K$ **in parallel do**
    $\theta_i^{t,im} \leftarrow$ ***Federated Cross-Correlation Learning*** $(Z_i, \overline{Z}, \theta_i^{t-1})$
    $\theta_i^t \leftarrow$ ***Federate Continual Learning*** $(\theta_i^*, \theta_i^{t-1}, \theta_i^{t,im})$
return θ_i^T

Federated Cross-Correlation Learning $(Z_i, \overline{Z}, \theta_i^{t-1})$
$\mathcal{M}_i \leftarrow (Z_i, \overline{Z})$ by Eq. (1)
$\mathcal{L}_i^{Col} \leftarrow (\mathcal{M}_i, \lambda_{Col})$ through Eq. (2)
$\theta_i^{t,im} \leftarrow \theta_i^{t-1} - \eta\nabla\mathcal{L}_i^{Col}$
return $\theta_i^{t,im}$ to i^{th} participant

Federated Continual Learning $(\theta_i^*, \theta_i^{t-1}, \theta_i^{i,im})$:
for $e = 1, 2, ..., E$ **do**
  $Z_{i,pvt}^{t,im} = f(\theta_i^{t,im}, X_i)$
  $\mathcal{L}_i^{CE} \leftarrow \mathtt{CE}(Z_{i,pvt}^{t,im}, Y_i)$ in Eq. (3)
  $\mathcal{L}_i^{Inter} \leftarrow \mathtt{KL}(Z_{i,pvt}^{t,im}, f(\theta_i^{t-1}, X_i))$ in Eq. (4)
  $\mathcal{L}_i^{Intra} \leftarrow \mathtt{KL}(Z_{i,pvt}^{t,im}, f(\theta_i^*, X_i))$ in Eq. (5)
  $\mathcal{L}_i^{Dual} = \mathcal{L}_i^{Inter} + \mathcal{L}_i^{Intra}$
  $\mathcal{L}_i^{Loc} = \mathcal{L}_i^{CE} + \lambda_{Loc}\,\mathcal{L}_i^{Dual}$
  $\theta_i^{t,im} \leftarrow \theta_i^{t,im} - \eta\nabla\mathcal{L}_i^{Loc}$
$\theta_i^t \leftarrow \theta_i^{t,im}$
return θ_i^t to i^{th} participant

The typical supervision loss in Eq. (3) and dual-domain knowledge distillation loss in Eq. (6) are complementary to each other. The former requires models to learn a discriminative representation that is meaningful for classification tasks, while the latter helps to regularize the model with soft and rich information in both intra and inter domain. Thus, the overall training target is:

$$\mathcal{L}_i^{Loc} = \mathcal{L}_i^{CE} + \lambda_{Loc}\,\mathcal{L}_i^{Dual}, \tag{7}$$

where $\lambda_{Loc} > 0$ is a coefficient. As shown in Fig. 4, the features learned by $\mathcal{L}_i^{Dual}$ is more compact and separated in both intra and inter domain by enjoying the advantage of both typical supervision loss and dual-domain knowledge distillation loss, models show better discriminative features, producing promising intra and inter domain performance.

3.3. Discussion and Limitation

We describe *FCCL* in Alg. 1. *FCCL* constructs cross-correlation matrix with the average logits output. Therefore, *FCCL* is applicable when there are a large size of participants in federated learning, attributed to that the computation complexity for server side is $\mathcal{O}(K)$. Besides, *Federated Cross-Correlation Learning* does operation on logits output regardless of the specific model structure. Thus, when participants share same model structure (model homogeneity), *FCCL* is still capable. Assuming that there is no data heterogeneity among distributed data, the first term of $\mathcal{L}_i^{Col}$ in Eq. (1) would be close to zero, but the second term still disassociates different dimensions on logits output. On this basis, *FCCL* is model agnostic method and able to handle different degree of domain shift. However, we also note limitation on the requirement of task consistency. For multi-task setting, logits output may not only have distinct dimensions, but also contain different meanings for same dimension. This limitation is also shared by related methods [38, 50, 74, 91].

4. Experiments

Data and Model. We extensively evaluate our method on two classification tasks (*e.g.*, *Digits* [27, 37, 62, 68] and *Office-Home* [82]) with three public data (*e.g.*, *Cifar-100* [35], *ImageNet* [69] and *Fashion-MNIST* [87]). Specifically, *Digits* task includes four domains (*i.e.*, *MNIST* (M), *USPS* (U), *SVHN* (SV) and *SYN* (SY)) with 10 categories. The *Office-Home* task also have four domains (*i.e.*, *Art* (A), *Clipart* (C), *Product* (P) and *Real World* (R)). Note that for both tasks, data acquired from different domains present domain shift (**data heterogeneity**).

For these two classification tasks, participants customize models that can be differ from differentiated backbones and classifiers (**model heterogeneity**). For experiments, we set the model as *ResNet* [23], *EfficientNet* [78], *MobileNet* [26] and *GoogLeNet* [76] for these four domains.

Comparison Methods. We compare our method, *FCCL* with state-of-the-art approaches including *FedDF* [50], *FML* [72], *FedMD* [38], *RCFL* [16] and *FedMatch* [28]. We also compare *SOLO*, where participant trains a model on private data without federated learning. Since specific experimental settings are not totally consistent, we retain key features of methods for comparison.

Evaluation Metrics. We report the standard metrics to measure the quality of methods: accuracy, which is defined as the number of samples that are paired divided by the number of samples. Specifically, for evaluation intra and inter domain performance, we define as following:

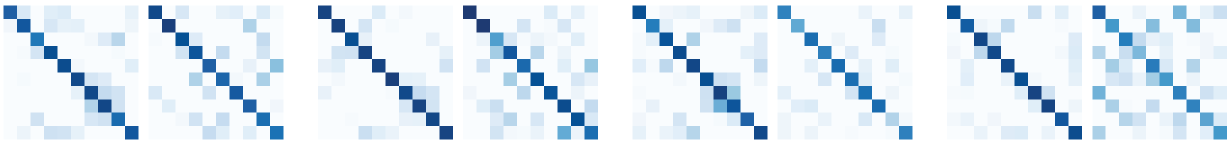
$$\mathcal{A}_i^{Intra} = \frac{\sum(\mathtt{argMax}(f(\theta_i, X_i^{Test})) == Y_i^{Test})}{|D_i^{Test}|}, \tag{8}$$

$$\mathcal{A}_i^{Inter} = \sum_{j\neq i}\frac{\sum(\mathtt{argMax}(f(\theta_i, X_j^{Test})) == Y_j^{Test})}{(K-1)\times|D_j^{Test}|}. \tag{9}$$

As for the method overall performance evaluation, we

Methods	Digits					Office-Home				
	$M\!\rightarrow$	$U\!\rightarrow$	$SV\!\rightarrow$	$SY\!\rightarrow$	AVG	$A\!\rightarrow$	$C\!\rightarrow$	$P\!\rightarrow$	$R\!\rightarrow$	AVG
SOLO	15.29	13.91	39.24	34.30	25.68	18.89	19.36	21.97	21.02	20.31
FedMD [38]	8.97	12.61	40.89	43.03	26.38	16.85	23.13	28.78	25.01	23.44
FML [72]	17.11	16.00	45.19	46.26	31.14	18.97	24.41	29.75	24.91	24.51
RCFL [16]	10.21	16.10	48.85	37.96	28.28	15.16	22.01	27.98	23.95	22.28
FedDF [50]	13.23	19.29	45.25	43.95	30.43	17.38	21.76	25.17	22.97	21.82
FedMatch [28]	9.22	14.76	46.28	36.05	26.58	19.05	25.24	28.73	24.35	24.34
FCCL	20.74	20.60	44.68	48.02	**33.51**	25.55	26.41	30.14	29.41	**27.88**

Table 1. **Comparison of inter domain performance with state-of-the-art methods.** $M\!\rightarrow$ means that private data is *MNIST* and respective model is tested on other domains in Eq. (9). AVG denotes average accuracy calculated from each domain. (The best average accuracy is marked in bold. The best entries in each domain are underlined. These notes are the same to others.)



(a) *MNIST* (b) *USPS* (c) *SVHN* (d) *SYN*

Figure 5. **Cross-correlation matrix visualization** for different domains on *Digits* task with *Cifar-100*. We visualize the cross-correlation matrix (Eq. (1)) with other models on public data (left) and private data (right) respectively. The left and right figure in each subfigure represent the cross-correlation matrix with other models on public data (*i.e.*, *Cifar-100*) and private data respectively. The matrix is 10×10. The darker the color, the closer the $\mathcal{M}_i^{uv}$ (Eq. (1)) is to 1.

adopt the average accuracy as metric. Besides, for these two classification tasks, *Digits* and *Office-Home* respectively contain 10 and 65 categories. Top-1 and Top-5 accuracy are adopted for these two tasks.

Implementation Details. In federated learning process, all participants adopt the same hyper-parameter setting (*i.e.*, $\lambda_{Col} = 0.0051$ like [91] and $\lambda_{Loc} = 1$). Models are trained using Adam optimizer [31] with batch size of 512 and learning rate as 0.001 in both collaborative updating and local updating for all approaches. In terms of data scale, in *Digits* task, *MNIST*, *USPS*, *SVHN* and *SYN* are assigned to four participants. The size of corresponding private data is set to 150, 80, 5000 and 1800 respectively. As for *Office-Home* task, each participant is individually assigned with *Art*, *Clipart*, *Product* and *Real World*, and the corresponding private data size is 1400, 2000, 2500, 2000. The number of unlabeled public data is 5000 for these two tasks. For pre-processing, we resize all input images into 32×32 with three channels for compatibility. We do communication for $T = 40$ rounds, where all approaches have little or no accuracy gain with more communications rounds. Besides, for *SOLO*, models are trained on private data for 50 epochs, which are also initial models for federated learning process.

4.1. Comparison with State-of-the-Art Methods

We provide comparison results with state-of-the-art methods on two image classification tasks (*i.e.*, *Digits* and *Office-Home*) with three public data (*i.e.*, *Cifar-100*, *ImageNet* and *Fashion-MNIST*).

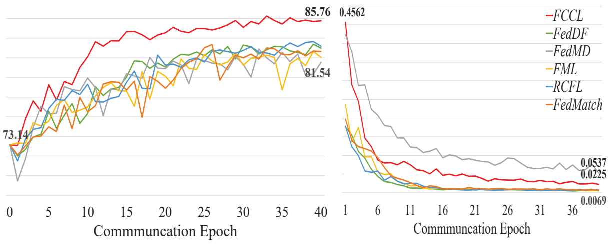
Inter Domain Analysis. We report the inter domain performance with state-of-the-art methods on Tab. 1. It clearly depict that under domain shift, *SOLO* present worst in these two tasks, demonstrating the benefits of federated learning. We observe that *FCCL* significantly outperforms better than counterparts. The Fig. 5 presents that *FCCL* achieves similar logits output between participants and redundancy within the logits output, confirming that *FCCL* successfully enforces the correlation of same dimensions and decorrelation of different dimensions on both public and private data.

Intra Domain Analysis. To compare the effectiveness of alleviating catastrophic forgetting, we show the intra domain performance in Tab. 2. Take the results of *Digits* task with *Cifar-100* as an example, our method outperforms the strong compared method, *RCFL*, by 2.30%. Besides, the intra domain accuracy via increasing communication rounds in Fig. 6a and optimization objective value in Fig. 6b present that *FCCL* suffers less periodic performance shock and is not prone to overfitting to current data distribution ($\overline{L}^{Loc} = 0.0225$), illustrating that *FCCL* is cable of balancing multiple knowledge, alleviating catastrophic forgetting.

Model Homogeneity Analysis. We further compare *FCCL* with other methods under model homogeneity. We set the shared model as *ResNet-18* and add the averaging parameters operation between *collaborative updating* and *local updating*. The Tab. 3 presents both inter and intra domain performance on *Office-Home* task with *Cifar-100*.

Methods	*Digits*				*Office-Home*			
	M	*U*	*SV*	*SY*	*A*	*C*	*P*	*R*
SOLO	70.20	74.19	74.57	73.60	65.27	60.50	74.68	54.28
FedMD [38]	77.30	80.05	77.73	87.72	66.17	60.63	76.35	56.60
FML [72]	80.66	79.75	78.58	88.87	81.46	65.58	79.82	65.07
RCFL [16]	82.59	81.05	78.79	91.40	65.13	61.33	76.44	55.78
FedDF [50]	82.95	78.84	78.46	91.30	66.10	60.44	75.70	55.98
FedMatch [28]	82.69	78.31	79.79	89.23	81.50	65.40	79.81	65.06
FCCL	88.84	84.42	78.55	91.23	81.51	65.42	79.84	65.16

Table 2. **Comparison of intra domain performance with state-of-the-art methods** on these two tasks with *Cifar-100*. The metric is evaluated on respective testing data in Eq. (8).



(a) Average Intra Domain Accuracy (b) $\overline{L}^{Loc}$

Figure 6. **Comparison of intra domain performance and optimization objective value in local updating** via increasing communication rounds on *Digits* task with *Cifar-100*.

Methods	Inter Domain				Intra Domain			
	$A\rightarrow$	$C\rightarrow$	$P\rightarrow$	$R\rightarrow$	*A*	*C*	*P*	*R*
SOLO	18.89	22.58	22.33	27.26	65.27	61.51	74.84	57.65
FedAVG [58]	57.85	54.05	55.72	60.18	66.71	60.90	74.29	57.49
FedMD [38]	61.03	62.41	62.45	62.55	66.50	61.75	73.63	58.10
FML [72]	39.56	36.94	32.73	42.00	74.87	60.73	77.19	60.71
RCFL [16]	61.52	59.56	57.56	63.59	67.16	61.39	73.33	58.58
FedDF [50]	61.10	57.92	62.19	60.41	66.69	60.69	74.12	57.69
FedMatch [28]	51.60	47.77	42.33	55.35	80.35	65.05	78.99	64.55
FCCL	64.48	62.33	63.26	64.86	81.38	65.47	79.40	65.19

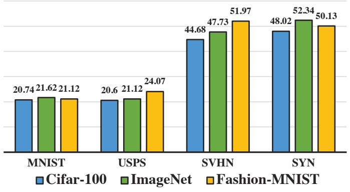
Table 3. **Comparison with state-of-the-art methods under model homogeneity** on *Office-Home* task with *Cifar-100*.

4.2. Diagnostic Experiments

To demonstrate how each component in *FCCL* contributes to overall performance, a series of ablation experiments are conducted. The proposed method, *FCCL* is comprised of two components: *Federated Cross-Correlation Learning* and *Federated Continual learning*.

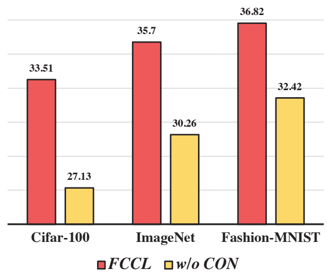
Federated Cross-Correlation Learning. To prove its robustness and stability, we evaluate the performance on different public data without label (*i.e.*, *Cifar-100*, *ImageNet* and *Fashion-MNIST*). The results in Fig. 7 suggest that *Federated Cross-Correlation Learning* achieves consistent performance in each domain. Moreover, it can be seen that it is more effective by the use of public data with rich categories (*ImageNet*) or simple detail (*Fashion-MNIST*).

Federated Continual Learning. We investigate the effectiveness of our core idea in handling catastrophic forgetting. As shown in Fig. 8, additionally considering dual-domain knowledge distillation (§ 3.2) leads to a substantial inter domain performance gain (*i.e.*, 6.38% on *Digits* task with *Cifar-100*), compared with w/o *CON* (optimization objective in local updating is only cross-entropy loss, L_i^{Ce} in Eq. (3)). In addition, the Fig. 8 illustrates that it also boosts the intra performance (*i.e.*, 3.88% with *ImageNet*). The Fig. 4 visualizes features in intra and inter domain cases. As seen, the proposed *Federated Continual Learning* begets a well discriminative feature space. This suggests that exploiting extra restriction signals in local updating is beneficial for alleviating catastrophic forgetting.



Public Data	AVG
Cifar-100	33.51
ImageNet	35.70
Fashion-MNIST	36.82

Figure 7. **Ablation study on *Federated Cross-Correlation Learning*** § 3.1 **with different public data for inter domain performance** on each domain performance (left) and overall performance (right) in *Digits* task.



Methods	AVG
Cifar-100	
FCCL	85.76
w/o *CON*	82.07
ImageNet	
FCCL	86.09
w/o *CON*	82.21
Fashion-MNIST	
FCCL	85.77
w/o *CON*	83.29

Figure 8. **Ablation study on *Federated Continual Learning*** § 3.2 **for inter (left) and intra (right) domain performance** on *Digits* task. w/o *CON* means that loss function is L_i^{CE} in Eq. (3).

5. Conclusion

This paper proposes a simple and effective method of *FCCL* for federated learning. *FCCL* is capable of handling heterogeneity problem and alleviating catastrophic forgetting. In particular, we construct cross-correlation matrix in collaborative updating to learn a generalizable representation. Meanwhile, we introduce knowledge distillation with inter and intra domain information in local updating, boosting inter and intra domain performance. Experimental results on classification tasks show that our method performs favorably in comparison with state-of-the-art methods.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China under Grants (62176188, 62141112, 41871243), the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, the Key Research and Development Program of Hubei Province (2021BAA187) and Zhejiang lab (NO.2022NF0AB01).

References

[1] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *CVPR*, pages 8218–8227, 2021. 3

[2] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, pages 535–541, 2006. 2, 3

[3] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. In *NeurIPS*, pages 15614–15624, 2020. 3

[4] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279*, 2019. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2, 3

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2, 3

[7] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 3

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 3

[9] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, pages 19–67, 2005. 3

[10] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, pages 1–1, 2021. 3

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2

[12] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *CVPR*, 2022. 2

[13] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Learning representations by predicting bags of visual words. In *CVPR*, 2021. 2

[14] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 3

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 3

[16] Gautham Krishna Gudur, Bala Shyamala Balaji, and Satheesh K Perepu. Resource-constrained federated learning with heterogeneous labels and models. In *ACM SIGKDD Workshop*, 2020. 6, 7, 8

[17] Xu Guo, Pengwei Xing, Siwei Feng, Boyang Li, and Chunyan Miao. Federated learning with diversified preference for humor recognition. In *IJCAI Workshop*, 2020. 1

[18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006. 2

[19] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018. 1

[20] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Fednas: Federated deep learning via neural architecture search. *arXiv preprint arXiv:2004.08546*, 2020. 2

[21] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large cnns at the edge. In *NeurIPS*, pages 14068–14080, 2020. 2

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 5

[25] Geoffrey E Hinton, Terrence Joseph Sejnowski, et al. *Unsupervised learning: foundations of neural computation*. MIT press, 1999. 2

[26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3, 6

[27] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE TPAMI*, pages 550–554, 1994. 2, 6

[28] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *ICLR*, 2021. 6, 7, 8

[29] Ce Ju, Ruihui Zhao, Jichao Sun, Xiguang Wei, Bo Zhao, Yang Liu, Hongshan Li, Tianjian Chen, Xinwei Zhang, Dashan Gao, et al. Privacy-preserving technology to help millions of people: Federated prediction model for stroke prevention. *arXiv preprint arXiv:2006.10517*, 2020. 1

[30] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 1, 2

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[32] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, pages 3521–3526, 2017. 3

[33] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016. 1

[34] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 1

[35] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 2, 6

[36] Solomon Kullback and Richard A Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, pages 79–86, 1951. 3

[37] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324, 1998. 2, 6

[38] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. In *NeurIPS Workshop*, 2019. 2, 3, 4, 5, 6, 7, 8

[39] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021. 1

[40] Qinbin Li, Bingsheng He, and Dawn Song. Model-agnostic round-optimal federated learning via knowledge transfer. *arXiv preprint arXiv:2010.01017*, 2020. 3

[41] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021. 2, 3

[42] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE SPM*, pages 50–60, 2020. 1

[43] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018. 1, 2

[44] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 3

[45] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *MedIA*, page 101765, 2020. 3

[46] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fed{bn}: Federated learning on non-{iid} features via local batch normalization. In *ICLR*, 2021. 1

[47] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, pages 2935–2947, 2017. 3

[48] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. In *NeurIPS Workshop*, 2020. 2, 3

[49] Renjie Liao, Alexander Schwing, Richard S Zemel, and Raquel Urtasun. Learning deep parsimonious representations. In *NeurIPS*, pages 5083–5091, 2016. 2

[50] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *NeurIPS*, pages 2351–2363, 2020. 2, 3, 4, 5, 6, 7, 8

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 3

[52] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021. 1, 3

[53] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *ICPR*, pages 2262–2268, 2018. 3

[54] 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 *AAAI*, pages 13172–13179, 2020. 1

[55] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, pages 7765–7773, 2018. 3

[56] Christopher May and Susan K Sell. *Intellectual property rights: A critical history*. Lynne Rienner Publishers Boulder, 2006. 2

[57] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, pages 109–165. Elsevier, 1989. 2, 3

[58] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017. 1, 2, 5, 8

[59] Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *IEEE SSP*, pages 19–38, 2017. 1

[60] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *PR*, pages 521–530, 2012. 1

[61] Pedro Morgado and Nuno Vasconcelos. Nettailor: Tuning the architecture, not just the weights. In *CVPR*, pages 3044–3054, 2019. 3

[62] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011. 2, 6

[63] Cuong V Nguyen, Alessandro Achille, Michael Lam, Tal Hassner, Vijay Mahadevan, and Stefano Soatto. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019. 3

[64] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, pages 1345–1359, 2009. 1

[65] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *ICLR*, 2020. 3

[66] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset Shift in Machine Learning*. Mit Press, 2009. 1

[67] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 3

[68] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018. 2, 6

[69] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, pages 211–252, 2015. 1, 2, 3, 6

[70] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[71] Felix Sattler, Arturo Marban, Roman Rischke, and Wojciech Samek. Communication-efficient federated distillation. *arXiv preprint arXiv:2012.00632*, 2020. 3

[72] Tao Shen, Jie Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Kun Kuang, Fei Wu, and Chao Wu. Federated mutual learning. *arXiv preprint arXiv:2006.16765*, 2020. 2, 3, 6, 7, 8

[73] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. In *NeurIPS Workshop*, 2019. 1, 2

[74] Lichao Sun and Lingjuan Lyu. Federated model distillation with noise-free differential privacy. *arXiv preprint arXiv:2009.05537*, 2020. 2, 5, 6

[75] Lichao Sun and Lingjuan Lyu. Data-free knowledge distillation for heterogeneous federated learning. In *ICML*, 2021. 3

[76] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 6

[77] Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In *NeurIPS*, pages 21394–21405, 2020. 1, 2

[78] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 3, 6

[79] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, 2021. 3

[80] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 3, 4

[81] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 3

[82] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 2, 6

[83] William T Vetterling, William H Press, Saul A Teukolsky, and Brian P Flannery. *Numerical Recipes: Example Book C (The Art of Scientific Computing)*. Press Syndicate of the University of Cambridge, 1992. 3

[84] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, page 3152676, 2017. 1

[85] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *CVPR*, pages 184–193, 2021. 3

[86] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *CVPR*, pages 10734–10742, 2019. 2

[87] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 2, 6

[88] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, and Chengqi Zhang. Multi-center federated learning. *arXiv preprint arXiv:2108.08647*, 2021. 2

[89] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM TIST*, pages 1–19, 2019. 1, 2

[90] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pages 6210–6219, 2019. 3

[91] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. 2, 3, 4, 6, 7

[92] Edvin Listo Zec, John Martinsson, Olof Mogren, Leon René Sütfeld, and Daniel Gillblad. Specialized federated learning using mixture of experts. *arXiv preprint arXiv:2010.02056*, 2020. 2

[93] 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. 3

[94] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *ICLR*, 2021. 2

[95] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 1