

FEDERATED FORGETTING-RESISTANT REPRESENTATION LEARNING

Hui Wang¹, Jie Sun^{2*}, Tianyu Wo³, Xudong Liu^{1,2}

¹ School of Computer Science and Engineering, Beihang University, Beijing, China

² Zhongguancun Laboratory, Beijing, China

³ SKLSDE, School of Software, Beihang University, Beijing, China

ABSTRACT

Continuous learning faces the challenge of catastrophic forgetting. Our research findings indicate that in unsupervised federated continual learning (UFCL), the limited model capacity and interference among participants are the key factors contributing to this problem. Specifically, the fixed capacity of the model restricts its ability to retain historical knowledge. Besides, the indiscriminate aggregation of weights from multiple participants can cause interference, damaging the model memory. To address these challenges, we propose FedFRR, a federated anti-forgetting representation learning approach. FedFRR fits the participants' data distribution through a weighted combination of primary network units (PNU) in the model and optimizes model memory by adjusting the structure of PNUs. Additionally, FedFRR addresses interference by truncating the PNU with less weight change, thus reducing the scope of weight aggregation. The experimental results demonstrate that FedFRR achieves state-of-the-art performance, significantly enhancing the model's anti-forgetting ability.

Index Terms— Federated Learning (FL), UFCL, Anti-forgetting, Weight truncation

1. INTRODUCTION

Federated Continual Learning is a novel paradigm that embeds Continual Learning [1] into the Federated Learning (FL) framework [2]. It allows a participant's model to continually learn from its local data as well as the accumulated knowledge acquired by other participants, which is highly compatible with scenarios such as Edge Computing (EC) and the Internet of Things (IoT). In practical EC and IoT scenarios, models often need to incorporate unsupervised learning techniques due to the scarcity of labeled training data, which poses an urgent need for unsupervised federated continual learning (UFCL). Nonetheless, there are still significant challenges in implementing UFCL in practice. One such challenge is catastrophic forgetting, which refers to the phenomenon where the model gradually shifts its learned parameters from previous data towards the new data during optimization, resulting in the loss of previously acquired knowledge [3]. Although

catastrophic forgetting has drawn some attention recently, the related research (e.g., [4, 5]) has primarily focused on the supervised learning scenario, leaving the problem under UFCL largely unexplored.

This paper focuses on the challenges posed by catastrophic forgetting in the realm of UFCL. To conduct a comprehensive study, we perform an experiment to validate the effect of catastrophic forgetting under a specific UFCL scenario. The detailed experimental setup can be found in the appendix [6]. The result indicates that forgetting can cause an accuracy drop of up to 38.9%. In fact, a fixed model structure leads to limited model capacity, and “free” parameters in the model tend to saturate as more data distributions are introduced from participants. This may be an essential reason for the decline in model memory. Additionally, we find that the gradient aggregation of participants in the FL may also interfere, accelerating the forgetting of the distributions that have already been fitted by the model.

In previous research [7], it has been suggested that incorporating assumptions about participants' data distribution can enhance the performance of the model. Building upon this understanding, we propose a novel approach that considers each participant's data distribution as a composite of multiple unidentified fundamental distributions (FD). Our approach further utilizes various primary network units (PNU) within the model framework to effectively model these FDs. By adjusting the structure and quantity of the PNUs with weight gradient truncation, we can directly optimize the performance of the model's memory. To the best of our knowledge, this paper is the first to focus on anti-forgetting in UFCL settings.

Contributions. (1) We conducted preliminary experiments, which revealed that forgetting can pose a threat to the performance of the UFCL model. (2) We propose a novel federated anti-forgetting representation learning method under UFCL settings, referred to as FedFRR. FedFRR utilizes distribution mixing and weight gradient truncation techniques to mitigate forgetting. (3) Through extensive experiments on both synthetic and real-world datasets, we demonstrate that FedFRR outperforms the state-of-the-art approaches. On average, FedFRR shows performance improvements of **7.8%** and **15.9%** in terms of model representation and anti-forgetting.

2. METHODOLOGY

2.1. Problem Formulation

FedFRR focuses on a specific UFCL scenario that contains a global coordinator and P participants who continuously process streaming unlabeled samples $\{x\}_1^\infty$. We divide the samples into different stages, the stage capacity recorded as C_p^s , the sample distribution of participant $p \in [1, \dots, P]$ in stage $s \in [1, \dots, S]$ is denoted as \mathcal{P}_p^s . To simplify the problem, we use a consistent stage capacity, i.e., $C_p^i = C_p^j, i, j \in [1, \dots, S]$. FedFRR uses contrastive learning [8] technique to extract the representation from participants' samples while improving the model anti-forgetting. Specifically, in stage s , let $\mathcal{M}(x, \theta_p^s): \mathcal{X} \rightarrow \mathcal{R}$ parameterized by θ_p^s is the model of participant p mapping the x to the representation r in \mathcal{R} . We denote by R_{cl} the expected contrastive loss of participant p in s stage $R_{cl} = \mathbb{E}_{x \sim \mathcal{P}_p^s} [L_{cl}(\mathcal{M}(x, \theta_p^s))]$ and R_{af} the expected anti-forgetting loss $R_{af} = \mathbb{E}_{x \sim \mathcal{P}_p^s, x' \sim \mathcal{P}_p^k} [\sum_{k < s} L_{af}(\mathcal{M}(x, \theta_p^s), \mathcal{M}(x', \theta_p^k))]$.

The optimization objective is formally defined as:

$$\arg \min_{\theta^S} \mathcal{L}(\theta^S) = \frac{1}{PS} \sum_{p=1}^P \sum_{s=1}^S \mathbb{E}_{x \sim \mathcal{P}_p^s, x' \sim \mathcal{P}_p^k} [L_{cl}(\mathcal{M}(x, \theta_p^s)) + \sum_{k < s} L_{af}(\mathcal{M}(x, \theta_p^s), \mathcal{M}(x', \theta_p^k))], \quad (1)$$

where L_{cl} represents the unsupervised contrastive loss, L_{af} encourages the model $\mathcal{M}(\cdot, \theta_p^s)$ of participant p in the current stage s to have a similar performance to $\mathcal{M}(\cdot, \theta_p^k)$ in any historical stage k , i.e., maintaining memories in historical stages.

2.2. Federated Model Architecture

FedFRR introduces a new formulation by assuming the distribution \mathcal{P}_p^s of the participant p in stage s is a weighted mixture of U fundamental distributions (FD) $\{\tilde{\mathcal{P}}_u\}_{u=1}^U$, i.e., $\mathcal{P}_p^s = \sum_{u=1}^U \lambda_u \tilde{\mathcal{P}}_u$, where the mixing coefficients λ_u stand for the probabilities of sample in \mathcal{P}_p^s coming from $\tilde{\mathcal{P}}_u$. Figure 1 shows the proposed model architecture. It comprises an input layer, an output layer, and multiple primary network units (PNU). All PNUs share the input, and the output is a weighted mixture of all PNUs' output. FedFRR does not predefine FDs but instead fits them through PNUs in optimization.

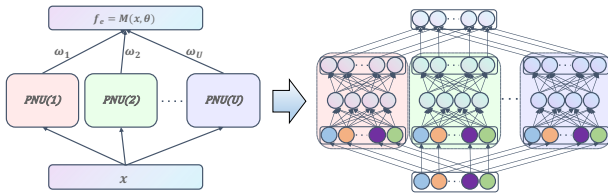


Fig. 1. Federated model architecture.

The multi-PNU architecture brings the following benefits: (1) The anti-forgetting ability of the model is positively correlated with its structure [3]. The more complex the structure (or more weights), the wider the distribution it

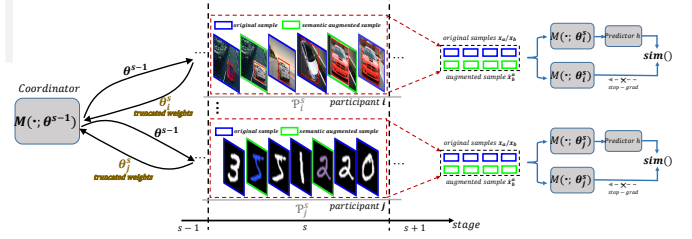


Fig. 2. FedFRR architecture. In stage s , the original and augmented sample pairs are processed by the model $\mathcal{M}(\cdot, \theta_p^s)$. Then a prediction MLP h is applied on one side, and a stop-gradient operation is applied on the other side. Model $\mathcal{M}(\cdot, \theta_p^s)$ maximizes the similarity between both sides.

can fit, i.e., the more knowledge it can learn. Compared to the Regularization-based [9], Rehearsal-based [4] methods, FedFRR can fundamentally improve the model memory by adding PNUs. (2) FedFRR can adjust the fitting ability of the PNU by changing its structure, thereby improving its representation performance. (3) FedFRR will truncate the PNU with small weight gradient changes in the optimization, reducing interference to the model memory.

2.3. Contrastive Optimization of the Federated Model

As shown in Figure 2, the coordinator distributes the model $\mathcal{M}(\cdot, \theta^{s-1})$ to all participants at the beginning of stage s . Based on the *SimSiam* [8] method, participant p uses the local samples $\{x\}_1^{C_p^s} \sim \mathcal{P}_p^s$ to perform contrastive optimization on the local model $\mathcal{M}(\cdot, \theta_p^{s-1})$, and then uploads the weight gradient $\Delta\theta = \theta_p^{s-1} - \eta \mathcal{L}$ to the coordinator. The coordinator aggregates the gradients using the formula 3.

$$\theta^s \leftarrow \theta^{s-1} + \sum_{p=1}^P \Phi(\theta_p^{s-1} - \eta \mathcal{L}), \theta^s \triangleq \cup \{\varphi_u^s \mid u \in [1, \dots, U]\}, \quad (3)$$

where η and \mathcal{L} denote the model learning rate and loss function. φ_u^s represents the weights of the u -th PNU in θ^s . $\Phi(\cdot)$ denotes an operation of zeroing the weights of the $T \in [0, \dots, U]$ PNUs with the slightest weight gradient change, i.e., weight truncation. FedFRR ignores the PNUs that have little effect on fitting the sample distribution through weight truncation, minimizing the scope of weight aggregation, as it may worsen the model memory. To accelerate the optimization, FedFRR uses semantic interpolation (see section 2.4) to augment the sample in stage. Therefore, a participant in stage s includes original and augmented samples x and \tilde{x} , and the following criteria should be met to ensure the invariance [10] of the representation learned by the model $\mathcal{M}(\cdot, \theta_p^s)$.

$$P(\mathcal{P}_p^s \mid \mathcal{M}(\tilde{x}_j^i, \theta_p^s), SIAug(x_i, x_j)) = P(\mathcal{P}_p^s \mid \mathcal{M}(x_i, \theta_p^s)) \cup P(\mathcal{P}_p^s \mid \mathcal{M}(x_j, \theta_p^s)), \forall x_i, x_j \sim \mathcal{P}_p^s, \quad (4)$$

where \tilde{x}_j^i represents the augmented sample from x_i, x_j through operation $SIAug(\cdot)$. FedFRR learns representations by minimizing the objective in Formula 2 over the samples x

$$\frac{1}{PS} \sum_{p=1}^P \sum_{s=1}^S \left[\sum_{i,j,k=1}^C -\log \frac{\exp(\text{sim}(\mathcal{M}(x_i, \theta_p^s), \mathcal{M}(\tilde{x}_j^i, \theta_p^s))/\tau)}{\mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathcal{M}(x_i, \theta_p^s), \mathcal{M}(x_k, \theta_p^s))/\tau) + \exp(\text{sim}(\mathcal{M}(x_i, \theta_p^s), \mathcal{M}(\tilde{x}_j^i, \theta_p^s))/\tau)} + \alpha \sum_{m < s} KL(\mathcal{M}(\cdot, \theta_p^s), \mathcal{M}(\cdot, \theta_p^m)) \right] \quad (2)$$

and \tilde{x} . P and S are the numbers of participants and stages. C is the stage capacity. $\text{sim}(\cdot)$ denotes the representation cosine similarity function, τ is the temperature and $\mathbb{1}_{[k \neq i]} = 1$ if $k \neq i$. α is the weighting of the forgetting penalty. $KL(\cdot)$ is the Kullback-Leibler divergence [11].

2.4. Semantic Interpolation Augmentation

The purer the distribution \mathcal{P}_p^s of participant p in stage s , i.e., the more consistent the essential features of the samples in \mathcal{P}_p^s are, the more beneficial it is to improve the fitting effect of PNU [7]. However, this may not be guaranteed effectively in an actual UFCL scenario. For instance, the images captured by a traffic camera at adjacent times may be vehicle and pedestrian. Therefore, FedFRR augments the samples in \mathcal{P}_p^s using the semantic interpolation [12] technique. For the image samples, semantic interpolation captures the areas containing important information (i.e., areas strongly related to essential features) in the original samples through a Semantic Percent Map (SPM). Then it mixes the areas by cut-and-paste at symmetrical locations. The augmented sample retains the original sample's essential features. The semantically augmented sample \tilde{x}_j^i from x_i and x_j can be expressed as:

$$\tilde{x}_j^i = (1 - BM_\lambda) \odot x_i + \Gamma(BM_\lambda \odot x_j), \quad (5)$$

where \odot denotes element-wise multiplication, the random value λ from a beta distribution $\text{Beta}(\beta, \beta)$. The BM_λ denotes a binary mask containing a random square region whose area ratio to the original sample is λ . $\Gamma(\cdot)$ is a function that transforms (e.g., rotate, grayscale change) the cutout region of x_i and x_j to increase the diversity of the augmented sample \tilde{x}_j^i . FedFRR cuts out the high semantic value region BM_λ from x_i and pastes it into the same region of x_j , and vice versa. This technique strengthens the attention paid to the essential feature in the original sample. It avoids the noise caused by traditional augmentation methods, such as Mixup [13]. To improve the semantic information in the BM_λ region, FedFRR uses the following Modified-SPM (MSPM) to measure each original image pixel's semantic relatedness to the essential feature. For a given participant's image sample $x \in \mathbb{R}^{d \times h \times w}$, we denote $FM_i^u(x)$ the i -th feature map in the last convolutional layer from the u -th PNU, and $\omega_u \in \mathbb{R}^d$ the u -th PNU weight corresponding to output feature in the Figure 1 (**Note:** Each element in the feature from the same PNU shares the weight connected to the output layer).

$$MSPM = \Psi \left(\sum_{i=1}^d \left(\frac{1}{U} \sum_{u=1}^U \omega_u FM_i^u(x) \right) \right), \quad (6)$$

where $\Psi(\cdot)$ denotes an operation that upsamples the feature map to match dimensions with sample x and normalizes it. Unlike the original SPM [12], MSPM does not rely on supervisory signals, making it more suitable for UFCL scenarios.

3. EXPERIMENTS

Datasets: We evaluate FedFRR using image classification datasets: MIXED [14], CMNIST [15], CCIFAR10 [16], FFHQ [17], and MiniImageNet [18]. **Baselines:** Our baselines consist of the supervised group (HLE [4], DER [5]), unsupervised group (SimSiam [8], RELIC [10]), and federated group (FedSimCLR [19], FedCA [20], FedWeIT [21]). **Implementation Details:** We use ResNet and multi-layer CNN for the PNU, the PNU number $U = 10$, participant number $P = 20$, the stage capacity $C = 100$, and the number of PNU with truncated weight $T = 0.1U$. In each stage, the augmented sample number $A = 0.5C$. The forgetting penalty coefficient $\alpha = 0.5$ in Formula 2, and the learning rate $\eta=0.01$. The layout of PNUs includes *SI*: all PNUs have the same structure. *MS*: The model's first and second half of PNUs are implemented in two different structures. *MA*: two different structures alternately implement all PNUs. The metric *Forgetting* = $\text{MAX}\{ACC - ACC', 0\}$, ACC is the model accuracy after pre-training, and ACC' is the real-time accuracy in different stages. More details about the datasets, baselines, and algorithm can be found in the appendix [6].

3.1. Quantitative Evaluation

Comparison of the Model Accuracy and Forgetting. The model accuracy and forgetting comparisons are shown in Table 1. In general, FedFRR outperforms the three baseline groups by 8.6%, 8.8%, and 6.1% on average accuracy over all datasets, while forgetting has an average decrease of 11.6%, 20.8%, and 15.8%. The comparison indicates that under UFCL settings, distribution fitting and reducing the weight aggregation scope are effective ways to improve model representation and anti-forgetting performance simultaneously.

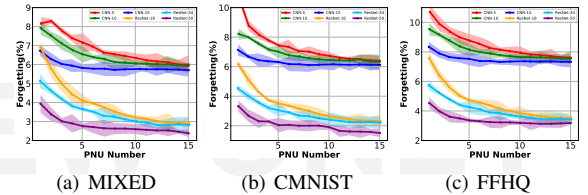


Fig. 3. The impact of PNU on model forgetting.

Evaluation on PNU Architecture. Figure 3 depicts the impact of the PNU architecture on model forgetting. For the ResNet-18/34/50 and CNN-5/10/15 PNUs, the more complex their structure, i.e., more parameters, the faster the forgetting decreases. Empirically, the ResNet or CNN PNUs may have a common forgetting lower bound. Figure 4 depicts the accuracy distribution of models composed of different PNUs. The accuracy on the three datasets increases by an average of 6.7%, 3.6%, and 7.3% as the PNU number from 1 to 3. The

Table 1. Comparison of the model accuracy and forgetting. We use the model composed of ResNet-18 PNUs.

Method	MIXED		CMNIST		CCIFAR10		FFHQ		MiniImageNet	
	Acc	Forgetting	Acc	Forgetting	Acc	Forgetting	Acc	Forgetting	Acc	Forgetting
HLE	80.41 \pm 0.59	10.30 \pm 0.84	86.27 \pm 0.39	09.12 \pm 0.64	43.98 \pm 0.91	06.14 \pm 0.12	59.48 \pm 0.76	08.20 \pm 0.62	30.18 \pm 0.26	06.20 \pm 0.74
DER	79.49 \pm 1.04	09.60 \pm 1.14	84.48 \pm 1.49	08.74 \pm 1.62	48.84 \pm 1.87	05.13 \pm 0.35	58.91 \pm 0.99	07.10 \pm 0.23	29.38 \pm 0.86	06.80 \pm 1.72
SimSiam	78.73 \pm 0.64	11.60 \pm 1.26	88.84 \pm 0.72	14.40 \pm 1.65	50.81 \pm 2.02	11.15 \pm 2.15	60.28 \pm 0.86	10.53 \pm 1.53	33.68 \pm 1.26	08.23 \pm 2.73
RELIC	80.83 \pm 1.38	12.30 \pm 0.94	87.47 \pm 0.86	13.01 \pm 0.23	49.21 \pm 1.12	09.14 \pm 1.43	61.72 \pm 0.16	09.21 \pm 0.72	32.74 \pm 0.73	09.11 \pm 1.12
FedSimCLR	82.65 \pm 2.04	13.20 \pm 1.47	89.57 \pm 0.61	11.12 \pm 2.64	52.11 \pm 1.32	08.61 \pm 0.92	61.18 \pm 1.76	11.33 \pm 1.42	34.28 \pm 0.66	10.22 \pm 2.11
FedCA	79.63 \pm 0.12	10.30 \pm 0.96	88.74 \pm 1.76	11.95 \pm 1.23	51.71 \pm 0.34	07.32 \pm 1.23	60.48 \pm 2.83	10.92 \pm 1.53	35.28 \pm 0.96	09.01 \pm 1.34
FedWeIT	83.29 \pm 1.39	07.20 \pm 0.84	90.27 \pm 1.03	07.14 \pm 0.72	53.23 \pm 1.87	04.82 \pm 1.13	64.88 \pm 0.96	05.73 \pm 0.42	36.82 \pm 1.76	05.11 \pm 0.13
FedFRR	86.79 \pm 1.24	04.90 \pm 1.12	91.62 \pm 0.71	04.22 \pm 0.17	54.78 \pm 0.37	04.01 \pm 0.31	66.08 \pm 0.46	05.43 \pm 0.92	39.52 \pm 0.16	03.32 \pm 0.97

difference in accuracy between models composed of ResNet and CNN PNU indicates that ResNet PNU is more conducive to improving the model’s representation ability. In addition, the accuracy distribution shows cohesion, obtaining a smaller Interquartile Range, indicating that the model’s representation performance is not sensitive to structures within the same group, which increases the implementation space of PNU.

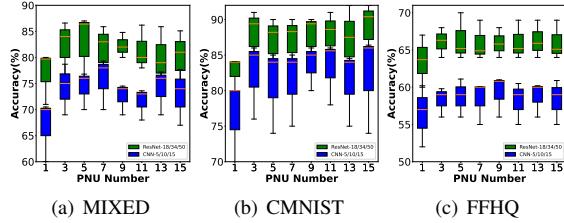


Fig. 4. The impact of PNU on model accuracy.

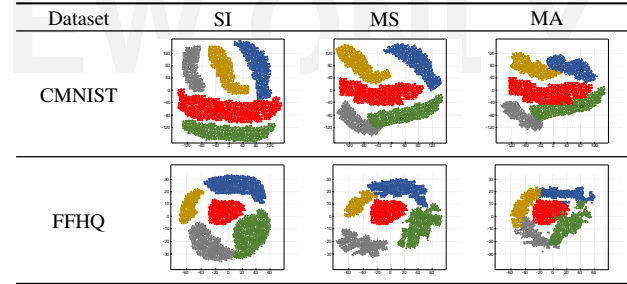
The Impact of Weight Truncation Ratio (WTR). Table 2 describes the impact of PNU layout and WTR on accuracy and forgetting. For different layouts, the accuracy is roughly the same, which means the accuracy is not sensitive to the PNU layout. With the increase of WTR, forgetting shows a decreasing trend and then increases, reaching its lowest at about WTR=0.1. The increase of WTR means the weight scope involved in aggregation decreases, reducing weight interference among participants. However, a further increase in WTR will result in the model losing more valuable weight information, leading to growth in the optimization stage.

Table 2. The impact of WTR on accuracy and forgetting.

Layout	WTR	MIXED			CMNIST		
		Stage	Acc	Forgetting	Stage	Acc	Forgetting
SI	0.0	293	85.21 \pm 1.91	07.12 \pm 0.72	154	90.17 \pm 1.74	10.41 \pm 1.01
	0.1	331	86.34 \pm 0.13	04.71 \pm 1.51	195	91.46 \pm 2.93	05.74 \pm 0.51
	0.3	403	85.43 \pm 1.81	08.71 \pm 1.63	236	91.47 \pm 1.72	07.37 \pm 1.51
	0.5	511	85.72 \pm 0.71	07.72 \pm 0.63	294	90.01 \pm 0.47	11.41 \pm 0.61
MS	0.0	302	86.31 \pm 1.64	09.12 \pm 1.71	172	89.71 \pm 1.64	09.41 \pm 2.72
	0.1	356	86.71 \pm 2.84	03.91 \pm 1.01	201	91.01 \pm 2.01	04.61 \pm 1.31
	0.3	420	85.11 \pm 0.26	06.21 \pm 1.41	224	90.45 \pm 1.31	06.51 \pm 2.31
	0.5	541	84.81 \pm 2.89	07.51 \pm 0.41	284	89.81 \pm 2.78	09.62 \pm 0.73
MA	0.0	318	84.61 \pm 0.12	07.61 \pm 0.63	162	90.71 \pm 1.41	09.74 \pm 0.27
	0.1	346	86.83 \pm 1.71	04.61 \pm 1.31	185	89.01 \pm 2.51	05.81 \pm 2.91
	0.3	415	85.81 \pm 1.61	05.64 \pm 0.21	229	91.11 \pm 0.13	08.76 \pm 2.34
	0.5	532	85.61 \pm 2.10	06.73 \pm 2.51	278	90.64 \pm 1.14	12.61 \pm 2.73

PNU Distribution Fitting Effect. Table 3 depicts the decomposition effect of PNU on sample distribution through feature projection (using *t*-SNE algorithm [22]). The layout *SI* represents the model composed of four CNN-10 PNUs, while *MS/MA* represent two CNN-10 and two ResNet-18 PNUs. The output features from different PNUs in the model exhibit

Table 3. The distribution fitting effect of PNU. Red represents the output features of the model, while other colors represent the output features of different PNUs in the model.



“high cohesion, low coupling”, indicating that PNUs effectively decompose and fit the participant samples. In the *MA* layout, there is a slight overlap between the feature distributions, indicating a poor decoupling effect, which may be caused by the alternating arrangement of PNUs in the model.

Table 4. Ablation results of sample augmentation methods.

Method	no augmen- tation	Vanilla	Mixup	SIAug(ours)
MIXED	78.71 \pm 1.21	80.12 \pm 0.43	82.61 \pm 1.91	85.84 \pm 0.71
CMNIST	84.61 \pm 0.51	87.51 \pm 1.43	86.71 \pm 1.63	91.01 \pm 1.43
CCIFAR10	44.01 \pm 2.72	47.71 \pm 2.71	50.02 \pm 1.82	53.74 \pm 1.34
FFHQ	57.51 \pm 0.71	62.62 \pm 1.94	60.71 \pm 2.41	64.71 \pm 1.84
MiniImageNet	34.32 \pm 2.51	36.91 \pm 1.42	37.47 \pm 1.72	38.91 \pm 2.03

Ablation Studies on Stage Sample Augmentation. Table 4 describes the impact of different augmentation methods on the model accuracy. Vanilla represents random rotation or adding Gaussian noise to the sample. Mixup and SIAug denote the interpolation [23] and semantic augmentation for the random sample pairs of participants. Results show that the semantic augmentation for the essential features of the sample achieves state-of-the-art performance on all datasets.

4. CONCLUSION

We first formulate the model anti-forgetting representation learning problem under the UFCL setting and propose FedFRR, based on distribution fitting and weight gradient truncation techniques. The experiments show that FedFRR effectively improves the performance of model representation and anti-forgetting, outperforming the baseline methods.

5. REFERENCES

- [1] 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,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [2] Brendan McMahan, Eider Moore, Daniel Ramage, and Hampson, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [3] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang, “Federated continual learning with weighted inter-client transfer,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 12073–12086.
- [4] Byung Hyun Lee, Okchul Jung, Jonghyun Choi, and Se Young Chun, “Online continual learning on hierarchical label expansion,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [5] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara, “Dark experience for general continual learning: a strong, simple baseline,” *Advances in neural information processing systems*, vol. 33, pp. 15920–15930, 2020.
- [6] Anonymous, “Appendix,” Website, 2023, <https://anonymous.4open.science/r/FedFRR-ICASSP2024/Appendix.pdf>.
- [7] Aoxiao Zhong, Hao He, Zhaolin Ren, and Quanzheng Li, “Feddar: Federated domain-aware representation learning,” *arXiv preprint arXiv:2209.04007*, 2022.
- [8] Xinlei Chen and Kaiming He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15750–15758.
- [9] Iman Mirzadeh and Razvan Pascanu, “Understanding the role of training regimes in continual learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 7308–7320, 2020.
- [10] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, and Buesing, “Representation learning via invariant causal mechanisms,” in *International Conference on Learning Representations*, 2020.
- [11] Solomon Kullback, *Information theory and statistics*, Courier Corporation, 1997.
- [12] Shaoli Huang and Xinchao Wang, “Snapmix: Semantically proportional mixing for augmenting fine-grained data,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 1628–1636.
- [13] Hongyi Zhang, Moustapha Cisse, and David Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *International Conference on Learning Representations*.
- [14] Hong-Wei Ng and Stefan Winkler, “A data-driven approach to cleaning large face datasets,” in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 343–347.
- [15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [16] Alex Krizhevsky, Geoffrey Hinton, et al., “Learning multiple layers of features from tiny images,” 2009.
- [17] Tero Karras and Samuli Laine, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [19] Ting Chen, Simon Kornblith, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [20] 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.
- [21] Jaehong Yoon, Wonyong Jeong, and Sung Ju Hwang, “Federated continual learning with weighted inter-client transfer,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 12073–12086.
- [22] Laurens Van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [23] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *International Conference on Learning Representations*, 2018.