

Relaxed Contrastive Learning for Robust Federated Models with Noisy Labels and Limited Clients

Girum Fitihamlak Ejigu, Apurba Adhikary, Choong Seon Hong*

Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea

{girumfitex, apurba, cshong} @khu.ac.kr

Abstract—Federated Learning enables collaborative model training across distributed clients without sharing raw data, making it suitable for privacy-sensitive applications. However, FL faces significant challenges due to noisy labels, which can degrade model performance, especially in non-IID environments. In this paper, we propose a robust federated learning method that combines symmetric cross-entropy loss with a modified relaxed contrastive loss to address noisy labels and improve the diversity of learned feature representations. We evaluate our approach on Fashion-MNIST, SVHN, and CIFAR-10 datasets under both symmetric and asymmetric noise settings. The results demonstrate that our method consistently outperforms multiple baselines, including FedAvg, FedProx, FedMixup, and FedLSR. Notably, our method shows strong performance even with low client participation rates, making it highly effective for scenarios where communication cost is a concern. This robustness and efficiency make our approach a valuable contribution to noise-robust federated learning.

Index Terms—Federated learning, contrastive learning, noisy label learning

I. INTRODUCTION

Federated Learning (FL) is a decentralized machine learning framework where clients collaboratively train a global model without sharing their local data [1]. Instead, only model updates are sent to a central server, aggregating them to improve the global model. This approach is advantageous for privacy-sensitive applications, as raw data remains on the clients' devices.

However, FL faces challenges from noisy labels, where incorrect or inconsistent labels in local datasets can significantly degrade model performance [2]. Noisy labels may arise due to human error or the varying conditions under which clients collect and label data [3]. This issue is further aggravated by the decentralized nature of FL, where the central server has limited visibility into local data quality.

Addressing noisy labels in FL is crucial, as they can distort local updates and harm the aggregated global model. Recent studies, [4], [5], have shown that even small fractions of noisy labels can disproportionately affect performance in non-IID

settings, making robust noise-handling strategies a priority in FL research. In recent works, several strategies have been proposed to handle noisy labels in the FL setting. For instance, FedLSR implements a local self regularization technique on clients to soften the effect of noisy labels. Fed-NCL, on the other hand, leverages a noise-aware contrastive loss to enhance local feature representations, making the learning process more robust to mislabeling [6]. FedCorr introduces a noise correction mechanism by filtering noisy updates at the client level before sending them to the server, improving the robustness of the global model to label errors [7].

Despite the advancements in noise-robust FL, achieving both communication efficiency and robustness to noise remains an open challenge. This is especially critical in non-IID settings, where data distributions vary across clients, and noisy labels can further complicate the aggregation of reliable updates. Therefore, there is a growing need for methods that not only enhance noise robustness but also operate efficiently under low client participation rates. Our proposed method addresses these dual challenges by incorporating a noise-robust loss function with a relaxed contrastive loss, demonstrating superior performance, particularly in scenarios where communication costs are high and client participation is limited. Through extensive experiments, we show that our approach outperforms several baselines, setting a new benchmark for noise-robust federated learning.

II. RELATED WORKS

A. Federated Learning

FL enables collaborative global model training while preserving client data privacy, making it ideal for privacy-sensitive applications like healthcare and finance. A key algorithm in FL is FedAvg [1], which averages local model parameters from clients. However, FedAvg struggles with non-IID data, where client data distributions vary. To address this, FedProx [8] adds a proximal term to the local objective, reducing client drift and improving convergence in non-IID settings. SCAFFOLD [9] further mitigates local bias by using control variates to correct local updates, improving convergence in heterogeneous data environments.

Beyond these methods, MOON [10] enhances FL by incorporating a model-contrastive loss that aligns local model representations with the global model, tackling non-IID data challenges. Similarly, [11] explores the viability of contrastive learning at the global model and client data scales on non-IID

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) and the National Research Foundation of Korea(NRF) (No. RS-2024-00352423) grant funded by the Korea government(MSIT) (No.2019-0-01287, Evolvable Deep Learning Model Generation Platform for Edge Computing), and supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023-00258649) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) *Dr. CS Hong is the corresponding author.

FL environment. Further improving on contrastive learning in FL, relaxed contrastive loss [12] addresses the issue of representation collapse, where feature diversity is lost in non-IID settings. The study proposed a divergence penalty to maintain diversity within classes while preventing collapsed representations. This method improves the transferability of learned features across clients, enhancing overall performance in decentralized training scenarios.

B. Federated Learning with Noisy Labels

Noisy labels present a significant challenge in FL because data is collected in a decentralized manner, which can lead to inconsistent or incorrect labels across clients. These noisy labels can arise from various factors, such as human annotation errors or differences in data quality across environments [2], [3]. Without proper handling, noisy labels can degrade global model performance as errors accumulate during aggregation. Several strategies have been suggested to address noisy labels in FL. Mixup [13], originally developed for centralized learning, generates virtual training samples by mixing data points and their labels, improving generalization. Another relevant approach is Co-teaching [14], which involves training two models simultaneously and allowing each model to select the cleanest samples from the other model, reducing the influence of noisy data during training. These methods help to filter out or smooth over noisy labels, improving model performance across clients.

Recent works have focused on identifying noisy clients and adjusting their contribution to the global model. FedLSR [15] introduces label smoothing regularization into the FL framework, which aims to prevent models from becoming overconfident in noisy labels. By applying a smoothing mechanism to the label distribution, FedLSR softens the impact of noisy labels, promoting more stable and generalized learning. FedNCL [6] suggests a mechanism for identifying noisy clients by evaluating both data quality and model divergence. It then applies robust, layerwise aggregation techniques that adaptively weight client updates, reducing the impact of noisy clients and addressing data heterogeneity. This approach ensures that cleaner, more reliable updates are prioritized during the global model aggregation process. Similarly, FedCorr [7] proposes a dynamic noise identification mechanism. By examining the dimensionality of the prediction subspaces on each client, FedCorr detects noisy clients and identifies incorrectly labeled samples based on per-sample losses. This method adjusts the aggregation process by down-weighting noisy clients and samples, thus ensuring that the global model remains robust against the influence of noisy data.

III. METHOD

A. Problem Definition

In an FL setting, let us consider N clients denoted as $\{C_1, C_2, \dots, C_N\}$ each with its own local dataset $\mathcal{D}_n = \{(x_i, y_i)\}$, where x_i are the input samples and y_i are the corresponding labels. The aim is to train a global model f_θ with parameters θ by collaboratively aggregating knowledge

from each client while preserving data privacy, as no client shares its raw data with the central server.

Each client C_n optimizes the local model by minimizing the empirical risk over its dataset, represented by the loss function $L_n(\theta)$. The server aggregates the local updates, typically using an averaging method like FedAvg [1] to update the global model θ . Mathematically, the optimization objective can be expressed as:

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N L_n(\theta), \quad (1)$$

However, in real-world FL settings, local datasets often contain noisy labels. These noisy labels may arise due to annotation errors, either from human annotators or automated labeling systems, and can vary their severity across clients. Noisy labels introduce significant challenges, as they lead to biased local updates, which, when aggregated deteriorate the performance of the global mode. In particular, we focus on non-IID data, where distribution of data and noise is not identical across clients, compounding the problem. To formalize this, let y_i^{true} represent the true label and y_i^{noisy} be the label observed by client C_n . The local loss function on client C_n is now computed on potentially noisy data;

$$L_n(\theta) = \mathbb{E}_{(x_i, y_i^{noisy}) \sim \mathcal{D}_n} [\ell(f_\theta(x_i), y_i^{noisy})] \quad (2)$$

Here, $\ell(\cdot)$ denotes the loss function, which is influenced by the noisy labels. To mitigate the negative impact of label noise, we combine a robust loss function called symmetric cross-entropy (SCE) [16] and a relaxed contrastive loss modified from [12] to enhance the diversity of feature representations and improve local updates in heterogeneous and communication efficient environment.

B. Client Update

In our FL framework, we address the problem of noisy labels in non-IID data environments with partial client participation by introducing a modified relaxed contrastive loss (RCL). This approach is inspired by the original work on contrastive learning from [12], which was designed for standard FL scenarios. We have tailored the RCL for robust FL, combining it with Symmetric Cross Entropy (SCE) [16] to reduce the adverse effects of inaccurate labels. This combination helps improve local updates by enhancing feature diversity and ensuring robust representation learning even in heterogeneous environments.

The SCE handles noisy label scenarios by balancing the impact of both the forward and backward cross-entropy losses, while the relaxed contrastive loss encourages feature separation within classes and alignment with the global model. Our proposed client update process is shown in Figure 1.

1) *Robust Relaxed Contrastive Loss (RCL) Process:* The RCL is designed to prevent representation collapse and encourage diverse feature learning during the local training phase of the client. The goal is to ensure that features from samples with the same (potentially noisy) label remain sufficiently distinct,

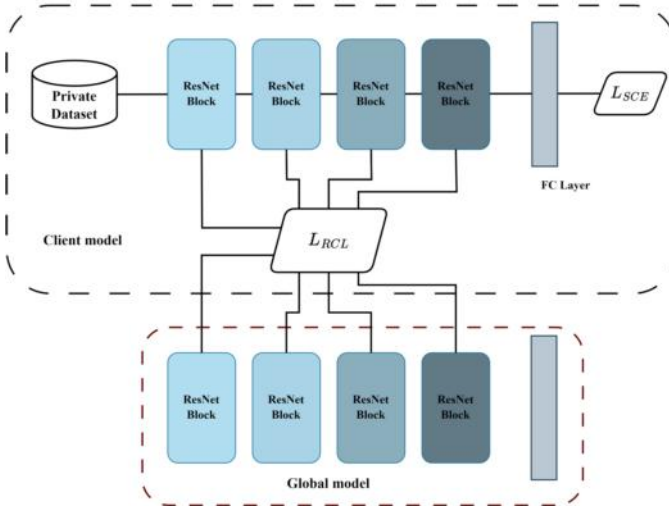


Figure 1. Client update process. The model utilizes SCE and RCL for mitigating noisy labels and enhancing feature alignment with the global model.

while also aligning local features with the global model to maintain consistency.

Intra-Class Contrastive Loss: Given a client C_n , the local model f_{θ_n} produces feature maps for each layer of the network. These feature maps, denoted by f_l^i represent the activations at layer l for client C_n . The local feature similarities are computed using cosine similarity between pairs of samples in the same class:

$$S(f_l^i, f_l^j) = \frac{f_l^i \cdot f_l^j}{\|f_l^i\| \|f_l^j\| \tau}. \quad (3)$$

where f_l^i and f_l^j are the feature vectors for samples i and j , respectively, and τ is the temperature parameter that controls how strongly positive and negative samples are separated. A mask M_{ij} is defined, where $M_{ij} = 1$ if samples i and j have the same label and $M_{ij} = 0$ otherwise. Then the intra-class contrastive loss that encourages samples with the same label to have higher similarity is as follows;

$$L_{intra} = -\log \frac{\sum_{j: M_{ij}=1} \exp(S(f_l^i, f_l^j))}{\sum_j \exp(S(f_l^i, f_l^j))} \quad (4)$$

Equation 4 enforces feature separation by ensuring that intra-class samples maintain sufficient diversity even under noisy labels.

Local-Global Contrastive Loss: to prevent that local client updates do not stray too far from the global model f_{θ} , we compute a local-global contrastive loss between the local features f_l^i and their corresponding global features f_l^g . This alignment is critical for keeping local models in sync with the global model across all clients. The local-global contrastive loss is given by;

$$L_{global} = -\log \frac{\sum_{j: M_{ij}=1} \exp(S(f_l^i, f_l^g))}{\sum_j \exp(S(f_l^i, f_l^g))} \quad (5)$$

By aligning local features with the global model, this loss ensures consistency in feature representations across all participating clients.

Divergence Penalty To prevent representation collapse, where intra-class features become overly similar, we add a divergence penalty following [12]. This term ensures that feature representations within the same class maintains a certain level of diversity. The divergence penalty is defined as:

$$L_{div} = \lambda \left(1 - \frac{1}{|P|} \sum_{i,j: M_{ij}=1} S(f_l^i, f_l^j) \right) \quad (6)$$

where λ is a hyperparameter that controls the strength of the penalty. This term discourages overly tight clustering of intra-class samples, especially in noisy environments. P refers to the set of positive pairs that belong to the same class as determined by the mask $M_{ij} = 1$.

The total RCL loss is calculated as the sum of the intra-class contrastive loss, local-global contrastive loss, and the divergence penalty across all layers:

$$L_{RCL} = \frac{1}{L} \sum_{l=1}^L (L_{intra}^l + L_{global}^l + L_{div}^l) \quad (7)$$

here L represents layers from a network model. This final loss encourages robust feature learning by promoting both intra-class separation and local-global alignment, while preventing feature collapse.

2) **Total Loss:** The total loss for the client update in our FL framework is a combination of the Symmetric Cross Entropy (SCE) loss, which is robust to noisy labels, and the relaxed contrastive loss (RCL), which enhances feature diversity and prevents representation collapse. This loss is given by;

$$L_{client} = L_{SCE} + \mu L_{RCL} \quad (8)$$

where μ controls the weight of RCL, allowing us to balance the focus between label noise robustness (SCE) and feature diversity (RCL), depending on the noise level and heterogeneity of the data.

Our method differs from the RCL in [12] in two key ways: first, we introduce a local-global contrastive loss to align local client features with the global model, ensuring consistency across clients and preventing local model drift. Second, we add an explicit divergence penalty to handle noisy labels, which prevents feature collapse caused by noisy data, while RCL focuses solely on intra-client feature diversity without targeting noise. These modifications make our approach more robust in FL environments with non-IID data and noisy labels.

C. Summary

Our framework tackles the challenges posed by noisy labels and non-IID data in FL settings with partial client participation. The global model is trained in a decentralized manner by aggregating updates from multiple clients, where each client performs local training on its private dataset. At each communication round, a subset of clients is selected, and the

server aggregates the local model updates using a weighted averaging strategy to update the global model.

On the client side, local training is conducted using a combination of SCE and RCL. The SCE loss provides robustness against noisy labels by balancing forward and backward cross-entropy terms, while RCL enhances feature diversity and mitigates representation collapse through intra-class contrastive learning, local-global alignment, and a divergence penalty. The total client loss is the weighted sum of SCE and RCL, with a parameter μ controlling the influence of RCL. Our framework, summarized in Algorithm 1, allows for effective learning even in heterogeneous and noisy environments, ensuring robust global model performance.

Algorithm 1 Robust Federated Learning with SCE and RCL

Server executes:

- 1: Initialize global model parameters θ_0
- 2: **for** each communication round $t = 1, 2, \dots, T$ **do**
- 3: Randomly select a subset of clients S_t
- 4: **for** each client $C_n \in S_t$ **in parallel do**
- 5: $\theta_{t+1}^n = \text{ClientUpdate}(\theta_t, n)$
- 6: **end for**
- 7: Aggregate local updates: $\theta_{t+1} = \frac{1}{|S_t|} \sum_{n \in S_t} \theta_{t+1}^n$
 {FedAvg}
- 8: **end for**

ClientUpdate(θ_t, n):

- 1: Receive global model $\theta_t^n \leftarrow \theta_t$
 - 2: **for** each local epoch $e = 1, 2, \dots, E$ **do**
 - 3: **for** each batch $(x_i, y_i^{\text{noisy}})$ in local dataset $\mathcal{D}_n$ **do**
 - 4: Compute model output $\hat{y}_i = f_{\theta_t^n}(x_i)$
 - 5: Compute SCE loss: $L_{\text{SCE}} = \text{SCE}(\hat{y}_i, y_i^{\text{noisy}})$
 - 6: Extract layer-wise local features $f_i^l \leftarrow f_{\theta_t^n}^l(x_i)$
 - 7: Extract global features $f_i^g \leftarrow f_{\theta_t}^g(x_i)$
 - 8: {Compute RCL loss for each layer:}
 - 9: $L_{\text{RCL}} = \frac{1}{L} \sum_{l=1}^L (L_{\text{intra}}^l + L_{\text{global}}^l + L_{\text{div}}^l)$
 - 10: Combine losses: $L_{\text{client}} = L_{\text{SCE}} + \mu L_{\text{RCL}}$
 - 11: Update local model: $\theta_t^n \leftarrow \theta_t^n - \eta \nabla_{\theta_t^n} L_{\text{client}}$
 - 12: **end for**
 - 13: **end for**
 - 14: **return** θ_t^n to the server
-

IV. EXPERIMENT

A. Datasets, Data Distribution and Non-IID Setup

We evaluate the performance of our method using three benchmark datasets: Fashion-MNIST [17], SVHN [18], and CIFAR-10 [19]. These datasets cover a range of complexity in terms of image classification tasks, allowing us to thoroughly assess the generalization and robustness of our approach under different noise conditions.

To simulate a realistic non-IID setting, we partition the data among clients using a Dirichlet distribution with a parameter of $\alpha = 0.5$. This approach allows for varying degrees of label distribution skew among the clients, creating more heterogeneous data partitions that mirror non-IID conditions

typical in FL environments. Each dataset is divided among 100 clients, but only a subset of 20 clients (20% participation) is selected for each round of FL training. This limited client participation further adds to the challenge by introducing additional variability in the training process.

B. Noise Generation and Baselines

We introduce label noise to simulate a noisy FL environment, using two types of noise generation techniques: symmetric noise and asymmetric noise [14]. Symmetric noise involves randomly flipping a percentage of labels to other classes. For our experiments, we evaluate noise levels at 15%, 30%, and 50%. On the other hand, asymmetric noise is designed to mimic label corruption where certain classes are confused with similar ones. We assess noise levels at 15%, 30%, and 45% for this type.

In our experiments, we compare the performance of our proposed method against several baseline approaches. FedAvg and FedProx are widely used and popular FL algorithms. FedAvg serves as the standard for aggregation in federated settings, while FedProx improves performance under non-IID conditions. In contrast, FedMixup, which implements the Mixup technique within an FL framework, is designed to be robust against noisy labels in centralized learning settings. FedLSR, on the other hand, incorporates a local self-regularization mechanism specifically designed to mitigate label noise in the FL environment. These methods provide a comprehensive range of techniques for comparison, highlighting both general FL strategies and those tailored to noisy environments.

C. Implementation Details

The FL process is conducted over 100 communication rounds, where each round involves five local training epochs. From a total of 100 clients, 20 clients (20% participation rate) are randomly selected for model updates in each round. The local model architecture is based on ResNet-10 [20], and for the computation of the RCL, we extract intermediate features from the first four ResNet block layers. This approach ensures a layer-wise enhancement of the feature representations during training.

For all datasets, the RCL parameter λ is set to 0.5, the temperature parameter τ is 0.1, and the divergence control parameter μ is also 0.1. Stochastic Gradient Descent (SGD) is employed as the optimizer with a learning rate of 0.01, momentum set to 0.9, and weight decay of 0.00001. All experiments are run on a single GeForce RTX 3080 GPU to ensure efficient training across the datasets.

D. Results

Table I summarizes the test accuracy comparison of our proposed method against several baseline approaches, FedAvg, FedProx, FedMixup, and FedLSR, under both symmetric and asymmetric noise conditions across three datasets: Fashion-MNIST, SVHN, and CIFAR-10. The experiments were conducted with varying noise levels and our method consistently

TABLE I
TEST ACCURACY (%) COMPARISON OF PROPOSED METHOD AND BASELINES UNDER MULTIPLE NOISE CONDITIONS. RESULTS ARE REPORTED AS AVERAGE AND STANDARD DEVIATION OF THE LAST FIVE ROUNDS.

Dataset	Method	Symmetric noise			Asymmetric noise		
		15%	30%	50%	15%	30%	45%
Fashion-MNIST	FedAvg	79.94±0.18	75.77±0.38	62.51±0.36	79.12±0.26	70.33±0.63	48.06±0.74
	FedProx	80.02±0.40	74.88±0.68	61.67±0.21	79.00±0.22	69.20±0.19	47.93±1.05
	FedMixup	80.41±0.17	78.11±0.41	74.82±0.48	79.23±0.11	74.07±0.32	53.44±1.36
	FedLSR	82.06±0.26	81.59±0.18	79.19±0.24	81.06±0.18	80.21±0.15	59.89±0.68
	Ours	82.24±0.22	81.47±0.37	78.76±0.14	81.13±0.17	78.33±0.44	60.25±2.51
SVHN	FedAvg	77.67±0.27	64.46±0.48	47.78±0.58	77.67±0.27	61.98±0.51	47.78±0.58
	FedProx	78.20±0.66	64.54±0.47	47.65±0.70	77.15±1.42	64.54±0.47	44.52±1.24
	FedMixup	76.90±1.21	69.22±0.53	57.09±0.57	73.31±1.77	58.98±1.78	42.05±2.42
	FedLSR	81.18±0.23	78.43±0.46	74.65±0.11	80.56±0.33	72.25±0.17	55.47±0.81
	Ours	86.64±2.03	79.97±0.50	76.16±0.46	83.82±1.94	75.48±0.38	53.76±0.87
CIFAR-10	FedAvg	55.75±0.46	45.28±0.43	31.80±0.24	55.37±0.20	45.13±0.35	34.11±0.23
	FedProx	56.57±0.27	45.46±0.38	30.69±0.61	56.17±0.42	44.97±0.29	34.13±0.23
	FedMixup	59.26±0.43	53.12±0.20	40.55±0.47	59.87±0.26	50.95±0.28	42.19±0.43
	FedLSR	54.96±0.17	54.61±0.24	46.44±0.27	54.98±0.16	54.58±0.30	48.42±0.45
	Ours	64.98±0.92	60.87±0.29	49.62±0.49	64.29±0.86	59.43±0.65	52.11±0.91

outperforms the baselines across the majority of settings, particularly under more severe noise conditions.

For Fashion-MNIST, under 50% symmetric noise, our approach achieves an accuracy of 78.16%, surpassing FedAvg, FedProx and FedMixup. However, FedLSR, an FL algorithm tailored for label noise resilience, exhibited competitive results with our approach, surpassing it in certain scenarios with 30% noisy labels. In the case of the SVHN benchmark, our method exhibits the highest accuracy score in almost all conditions, outperforming FedLSR and other baseline methods. Notably, under 15% symmetric noise, our method achieves 86.64%, showing much higher resilience than other baselines.

For the CIFAR-10 dataset, which presents a more challenging scenario, our method demonstrates clear improvements under high noise levels. For instance, under 50% symmetric noise, our method maintains an accuracy of 49.62%, outperforming FedAvg (31.80%) and FedLSR (46.44%). Similarly, under 45% asymmetric noise, our approach achieves a higher accuracy (52.11%) than all other baselines, showcasing the effectiveness of combining symmetric cross-entropy and RCL in improving robustness to noisy labels in FL environments. This consistent improvement across datasets and noise levels indicates the robustness of our method, particularly in challenging FL scenarios with both high noise levels and non-IID data.

We further investigate performance of each method across 100 communication rounds in Figure 2 under the SVHN dataset with a 15% symmetric noise. Our method consistently outperforms all baselines, showing rapid convergence than FedLSR and FedMixup, and achieving higher accuracy in the end. FedAvg and FedProx show comparable performance and faster convergence at earlier stage, with FedMixup initially outperforming FedLSR but eventually plateauing at a lower

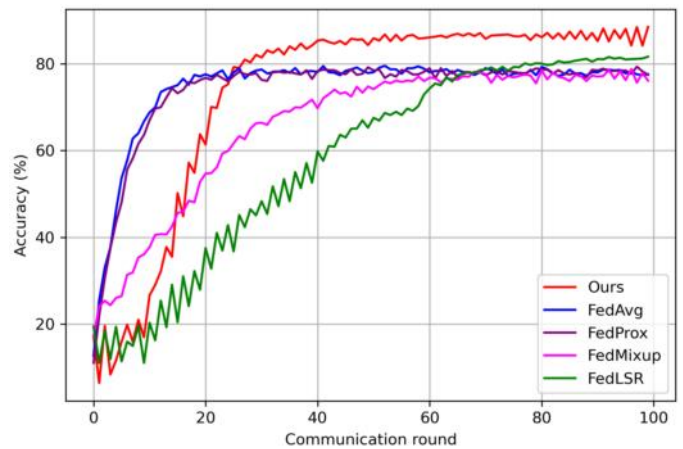


Figure 2. Test accuracy (%) of all methods over 100 communication rounds under 15% symmetric noise for the SVHN dataset.

TABLE II
TEST ACCURACY'S ON SVHN DATASET ON 15% SYMMETRIC NOISE AND NON-IID SETTING OF $Dir(0.5)$ WITH LIMITED CLIENT PARTICIPATION.

Method	10% client participation	5% client participation
FedAvg	65.41±0.51	53.36±1.70
FedProx	64.69±0.81	51.32±0.73
FedMixup	65.27±0.78	45.17±3.79
FedLSR	57.58±0.60	15.94±0.01
Ours	73.73±0.86	63.66±1.24

accuracy. FedLSR demonstrates a slower learning curve and the least stable performance, with significant fluctuations over the rounds. Notably, except for our method and FedLSR, the other approaches exhibit signs of overfitting as their accuracy stagnates in later rounds.

We perform additional experiments to test the robustness

of our proposed method under low client participation, a challenging scenario in FL. Using the SVHN dataset with 15% symmetric noise and a non-IID setting determined by $Dir(0.5)$, we compare test accuracies under two different client participation rates of 10% and 5%. The results are summarized in Table II, where we observe that performance degrades as client participation decreases, but our method consistently outperforms all baselines.

From the results, our method achieves the highest accuracy of 73.73% and 63.66% under 10% and 5% participation, respectively. Other methods, such as FedAvg and FedProx, exhibit lower accuracy, with FedMixup struggling significantly at 5% participation (45.17%). FedLSR, while designed to handle noisy labels, shows a sharp decline in accuracy, particularly at 5% client participation, suggesting that its robustness to low participation is limited. These results highlight the ability of our method to maintain superior performance even under extreme participation constraints.

The strong performance of our method under low participation rates, as observed in Table II, makes it particularly suitable for scenarios where communication efficiency is a critical concern. In FL settings, where only a fraction of clients may participate in each communication round due to high costs, our method demonstrates superior robustness and accuracy compared to other baselines. This advantage highlights its potential for deployment in resource-constrained environments, where minimizing communication overhead while maintaining model performance is essential.

V. CONCLUSION

In this work, we presented a robust federated learning approach that effectively addresses the challenge of noisy labels in decentralized and non-IID environments. By integrating a symmetric cross-entropy loss with a relaxed contrastive loss, our method enhances the diversity of feature representations while maintaining noise robustness. Extensive experiments performed on Fashion-MNIST, SVHN, and CIFAR-10 datasets demonstrate that our method consistently outperforms baseline approaches, including FedAvg, FedProx, FedMixup, and FedLSR, across various noise levels and settings.

Notably, our method excels under low client participation rates, making it well-suited for scenarios with high communication costs, where selecting only a fraction of clients per round is necessary. This efficiency in communication, combined with superior accuracy in face of noisy information, highlights the practical applicability of our approach in real-world FL applications. Future work could explore extending the approach to other model architectures and investigating its effectiveness in more complex real-world datasets.

REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [2] X. Fang and M. Ye, "Robust federated learning with noisy and heterogeneous clients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10072–10081.
- [3] Y. Chen, X. Yang, X. Qin, H. Yu, P. Chan, and Z. Shen, "Dealing with label quality disparity in federated learning," *Federated Learning: Privacy and Incentive*, pp. 108–121, 2020.
- [4] K. Kuo, P. Thaker, M. Khodak, J. Nguyen, D. Jiang, A. Talwalkar, and V. Smith, "On noisy evaluation in federated hyperparameter tuning," *Proceedings of Machine Learning and Systems*, vol. 5, pp. 127–144, 2023.
- [5] J. Li, J. Pei, and H. Huang, "Communication-efficient robust federated learning with noisy labels," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 914–924.
- [6] K. Tam, L. Li, B. Han, C. Xu, and H. Fu, "Federated noisy client learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [7] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 184–10 193.
- [8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.
- [9] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.
- [10] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 713–10 722.
- [11] G. F. Ejigu, Y. L. Tun, A. Adhikary, S. M. Kang, and C. S. Hong, "Effectiveness of model and data scale contrastive learning in non-iid federated learning," in *2023 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2023, pp. 301–304.
- [12] S. Seo, J. Kim, G. Kim, and B. Han, "Relaxed contrastive learning for federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 279–12 288.
- [13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
- [14] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.
- [15] X. Jiang, S. Sun, Y. Wang, and M. Liu, "Towards federated learning against noisy labels via local self-regularization," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 862–873.
- [16] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 322–330.
- [17] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
- [18] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 2. Granada, 2011, p. 4.
- [19] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis, University of Tront*, 2009.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.