Improving Model-Contrastive Federated Learning under Non-IID Data Constraints

Hao Pi

{ P23000410@njupt.edu.cn }

Portland Institute, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China

Abstract. To address the performance degradation and convergence instability issues caused by the non-IID nature of client data in federated learning, this paper proposes an enhanced model-contrastive federated learning framework and optimizes the existing methods at the client-side local training stage by introducing the dynamic scheduling strategy for the contrastive loss weight, temperature parameter annealing mechanism and multi-negative sample comparison strategy, which improves classification accuracy and achieves a better trade-off between supervised learning and contrast constraints in different training stages. For MINST, the proposed method achieves a classification accuracy of 95.18%, outperforming FedAvg and MOON by 0.19% and 0.11%, respectively. For Fashion-MNIST, the classification accuracy reached 81.87%, outperforming FedAvg and MOON by 1.42% and 1.19%, respectively. The experimental results demonstrate the effectiveness and feasibility of the proposed improved strategy under non-IID settings.

Keywords: Federated learning; Model-contrastive learning; Non-IID data; MOON.

1 Introduction

With the exponential increase in the amount of information today, data security is becoming increasingly important, and protecting data privacy has become an unavoidable topic. Companies and organizations are actively collecting large-scale data resources. With the progress of science and technology, how to obtain key parameters while protecting data security has become a problem that has been widely studied in recent years [1, 2].

The proposal of federated learning provides a feasible scheme for data privacy protection. Compared with traditional machine learning, it introduces a distributed cooperative training mechanism, which enables global optimization without requiring local data to be shared [3]. In recent years, many approaches have been proposed, resulting in significant advancements in communication efficiency, data security and privacy protection [4]. However, real-world communication environments are far more complex than experimental settings. Not only are there huge differences in computing power and communication conditions of client devices,

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365159

but also the data in the real world is mostly non-IID, which may lead to problems such as performance degradation, unstable training and model drift during the training process [5-8].

To solve the problem, this study builds an improved client training strategy based on the MOON algorithm, and realizes the phased optimization of the model training process by introducing dynamic scheduling of the contrastive loss weight, the annealing mechanism of temperature parameters and the comparison mechanism of multiple negative samples to improve training stability and learning accuracy [9]. Finally, the proposed method is tested on two public datasets to verify that it can achieve better classification performance than the existing methods, demonstrating the effectiveness and feasibility of the proposed method.

2 Method

2.1 Overall structure

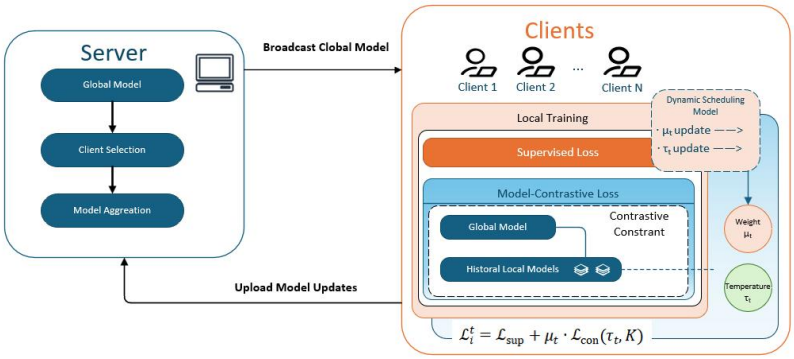

Fig.1. Framework of the improved MOON-based approach (Picture credit: Original)

This paper adopts a standard distributed training architecture (Figure 1), consisting of a server and diverse clients. In each communication round, the server and clients only exchange model parameters to achieve collaborative training without disclosing the data privacy.

During each round of federated learning, the server obtains local training parameters from clients, and broadcasts the current global model to each client after aggregation.

The client will train based on local data after receiving the global model. Different from the traditional federated learning algorithms which only depend on the supervised learning loss, this paper introduces the contrastive learning mechanism in the local training stage to reduce the interference of non-IID data on the model. In model contrastive learning, the current local model is not only aligned with the global model, but also the historical local models are introduced as a comparison reference. The historical local models come from the model state saved during the previous rounds of training on the client side to provide more stable representation alignment. After the optimization of the supervised loss and the model

contrastive loss, the two are jointly updated by means of weighted combination. Finally, the updated model parameters are uploaded to the server to complete a round of training.

Without changing the federated learning communication process and server-side aggregation strategy, this paper improves the training stability and performance of the model in non-IID settings through enhancing the local training objective on the client side and introducing dynamic contrastive constraints.

2.2 Comparison method settings

During local training on the client side, this paper constrains the consistency of model representations to guide the model to learn more stable feature representations under different training rounds and data distributions.

To address challenges in real-world scenarios, this paper integrates the aggregated global model together with several previously saved client models during the client optimization phase, so as to make the update of the current model more stable.

As for the workflow, upon receiving the global model from the server, each client conducts a forward pass with its local dataset, and extracts the corresponding feature representations. Then, the current model is compared with the global model and historical local models, respectively. By constructing multiple dynamic contrastive constraints, the current model is guided to maintain global consistency and local consistency during the optimization process.

The model contrastive loss and supervised loss are jointly optimized in the local training stage. The supervised loss ensures the prediction ability of the model on the local task, and the improved model contrast constraint is used to stabilize the training process of the model. In this way, the client can obtain a more stable training process under non-IID data, and provide more consistent model updates for the subsequent global model aggregation.

2.3 Learning strategy optimization

In order to improve the effect and stability of model contrastive learning under non-IID data, this paper improves the training strategy on the client side from three aspects: loss weight adjustment, temperature parameter control and comparative sample construction.

Firstly, this paper adjusts the contrastive loss weight μ and introduces the dynamic scheduling strategy for the contrastive loss weight μ_t. Because the weight of contrastive loss in the overall training objective has a significant impact on training performance, if it is set too large at the beginning, the model is prone to pay too much attention to alignment in the early stage of training, resulting in weakened supervised learning. Conversely, if the weight is too small, the contrastive constraint will not function effectively, and the model will still be affected by non-IID data [10]. Therefore, the proposed dynamic adjustment strategy sets μ_t to a small value at the initial stage of training to ensure that the model can fully learn the discriminative features of local data. With the increase in training rounds, μ_t gradually increases, allowing the contrastive constraint to provide stronger regularization in the middle and later phases of training. In order to realize the dynamic adjustment of the contrastive loss weight, μ_t is defined as follows:

$$\mu_t = \mu \cdot \min\left(1, \frac{t}{R_\mu}\right) \tag{1}$$

where μ_t is the weight of the contrastive loss at round t, μ is the maximum contrastive loss weight, t denotes the current communication round, R_μ is the weight scheduling parameter. When t<R_μ, μ_t increases linearly with the number of training rounds. When t≥R_μ, μ_t remains equal to μ.

Secondly, the annealing mechanism of the temperature parameter τ is adopted in this paper, so that the value of τ is gradually adjusted in the training process. The temperature parameter is used to control the smoothness of the similarity distribution in contrastive learning [11]. In real-world scenarios, model representations across clients may vary significantly. A fixed temperature parameter may therefore lead to large gradient fluctuations, which may affect training performance and stability. In this paper, a relatively large τ value is used during the initial training phase to weaken the sensitivity of representation differences to the loss function and avoid excessive gradient fluctuations. As the number of training rounds increases, τ is gradually reduced, enabling the model to more clearly distinguish positive and negative samples. In order to improve the numerical stability of gradient updates in the process of contrastive learning, the update rule of τ_t is defined as follows:

$$\tau_t = \tau_{start} + (\tau_{end} - \tau_{start}) \cdot \frac{t}{T} \tag{2}$$

$$\tau_{\text{start}} > \tau_{\text{end}}, \quad t = 1,2,\ldots,T \tag{3}$$

where τ_t denotes the temperature parameter at round t, τ_{start} and τ_{end} denote the temperature values at the beginning and end of training, respectively, satisfying $\tau_{start} > \tau_{end}$, T represents the total number of communication rounds.

Finally, this paper introduces the mechanism of multiple negative samples. The mechanism takes multiple historical versions of the local model together as negative samples to participate in contrastive learning. In the traditional contrastive learning methods, the construction of negative samples is generally limited, which makes it difficult to accurately capture the differences between models in different training stages and under different distribution conditions. This paper expands the negative sample set to incorporate more historical information into the training of the current model, enabling more consistent and accurate training outcomes. The mechanism can also effectively reduce the model oscillation across communication rounds. In this experiment, the contrastive loss is defined as:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp\left(\frac{\text{sim}\left(z_i^t, z_g^t\right)}{\tau_t}\right)}{\exp\left(\frac{\text{sim}\left(z_i^t, z_g^t\right)}{\tau_t}\right) + \sum_{k=1}^{K} \exp\left(\frac{\text{sim}\left(z_i^t, z_i^{t-k}\right)}{\tau_t}\right)} \tag{4}$$

where z_i^t represents the feature representation extracted by the current local model at client i in the t-th round of training, z_g^t corresponds to that extracted by the server's global model; z_i^{t-k} represents the feature representation corresponding to the historical local model saved by the

client in the previous K rounds of training. K represents the number of historical models involved in the comparison.

In summary, the client jointly optimizes the supervised loss and the contrastive loss in the local training stage, and its overall training objective is defined as:

$$\mathcal{L}_i^t = \mathcal{L}_{\text{sup}} + \mu_t \cdot \mathcal{L}_{\text{con}}(\tau_t, K) \tag{5}$$

3 Experiment

3.1 Datasets

In this paper, two commonly used image classification datasets MNIST and Fashion-MNIST are used to verify the proposed method. In order to simulate the heterogeneity of data in the real federated learning scenario, the dataset is partitioned under a non-IID setting and the data is divided and allocated to 20 clients. There are significant differences in the class distributions across clients. Each client only uses its local data for model training. The training and test sets are divided in the same way as the original dataset, and the test set is only used to evaluate the performance of the global model and is not involved in raining.

3.2 Experimental environment

The experiment was completed in a GPU-accelerated computing environment based on Python. The server functions as the central unit for global model aggregation and experimental control, and the training process of the client is realized by simulation on the same computing platform. The experiment is implemented under the Pytorch framework, and is extended and improved based on the federated learning experiment framework. The federated learning process is consistent with the existing federated learning experimental framework, which ensures the reproducibility of the experimental process. For fair experimental comparison, all methods in this experiment adopt the same model structure and training configuration. The training process utilizes SGD as the optimizer, with the learning rate and momentum set to 0.01. and a momentum coefficient of 0.9. Local training involves 5 epochs and a batch size of 64.

3.3 Experimental results and analysis

Table 1. Accuracy Results (%) on on MNIST and Fashion-MNIST

Dataset / Method	FedAvg	MOON	MOON-Improve
MNIST (50 rounds)	94.99	95.07	95.18
Fashion-MNIST (100 rounds)	80.45	80.68	81.87

Table 2. Ablation Study on MNIST and Fashion-MNIST

Strategy	MNIST	Fashion-MNIST
MOON	95.07	80.68
+ μ scheduling	95.11	81.05
+ τ annealing	95.10	80.98

+ multi-negative	95.14	81.42
+ all strategies	95.18	81.87

Table 1 presents the predictive accuracy of various techniques on the MNIST and the Fashion-MNIST datasets. On both datasets, FedAvg, as the baseline method, achieves relatively lower performance compared with the other approaches.

Table 2 presents the ablation results of each improvement strategy. Each strategy contributes to performance improvement, and the multi-negative-sample mechanism yields the most significant gain. When the three strategies are combined, the highest accuracy is achieved, which demonstrates that the proposed modules are complementary.

In contrast, the MOON method with a model-contrastive learning mechanism has achieved a certain degree of performance improvement on both datasets. MOON-Improve achieves higher accuracy than the original MOON method on both MNIST and Fashion-MNIST. For Fashion-MNIST, the performance of the improved method is more evident, indicating that the training strategy proposed in this paper has better adaptability under more complex data distributions. The above results verify the effectiveness of the proposed method under non-IID settings.

3.4 Discussion

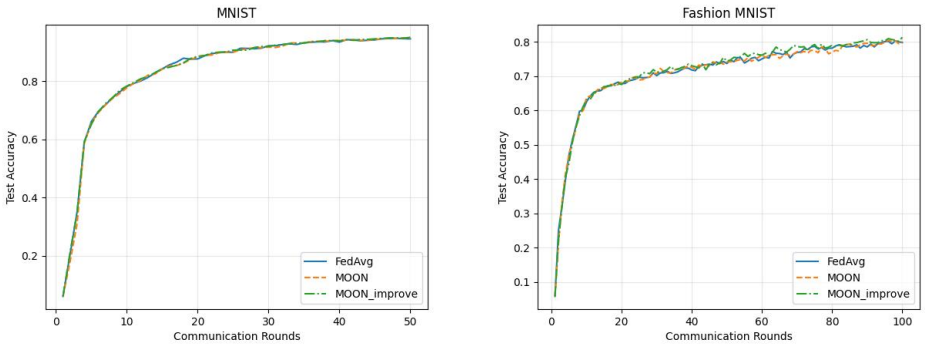

Fig.2. Test Accuracy versus Communication Rounds on MNIST and Fashion-MNIST (Picture credit: Original)

Figure 2 shows the trend of test accuracy with the number of communication rounds during the training process of different methods. From the experimental results of the MNIST dataset, each method has a faster convergence speed in the initial stage of training and reaches a stable performance level within a similar number of communication rounds.

On the Fashion-MNIST dataset, the performance difference between the above methods is more obvious. MOON-improve achieves relatively higher test accuracy toward the end of training, which shows that the improved strategy proposed in this paper provides a comparative performance advantage under more complex data distributions.

From the results shown in the table and the training curves, the proposed method effectively improves the final classification performance without significantly affecting the convergence speed. This result is consistent with the design objective, which shows that by refining the training approach on the client side, the model performance can be further improved under non-IID conditions.

4 Conclusions

This paper studies the improved training strategy based on model-contrastive learning under non-IID data conditions. Within the existing federated learning framework, the local training stage at the client side is enhanced by introducing dynamic contrastive loss weight scheduling, temperature parameter annealing mechanism and multi-negative sample contrastive strategy, the stability and accuracy of model contrastive learning are improved.

Compared with the FedAvg method and the the original MOON method, experimental results show that the improved strategy proposed in this paper achieves better classification accuracy on both datasets. Therefore, by reasonably designing the contrastive learning strategy during training, the federated learning model's overall behavior in non-IID scenarios can be effectively enhanced without significantly increasing the system complexity.

Of course, the experiment in this paper still have some limitations. For example, current experiments mainly focus on image classification tasks, and have not been validated on more complex model architectures or application scenarios. In addition, the multi-negative sample mechanism not only improves the performance but also increases computational and storage overhead. How to realize the trade-off between performance improvement and resource consumption is still worth further study.

Therefore, the future work can be carried out from multiple aspects, including applying the proposed training strategy to more complex real-world scenarios and exploring its feasibility. The proposed method also needs to be combined with other federated learning optimization techniques to further improve the model performance and training efficiency.

References

[1] Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology. pp. 1-19 (2019)
[2] Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine. pp. 50-60 (2020)
[3] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-Efficient Learning of Deep Networks from Decentralized Data. Proc. AISTATS. pp. 1273-1282 (2017)
[4] Kairouz, P., et al.: Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning. pp. 1-210 (2021)
[5] Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the Convergence of FedAvg on Non-IID Data. Proc. ICLR Workshop. (2020)
[6] Hsieh, K., Phanishayee, A., Mutlu, O., Gibbons, P.B.: The Non-IID Data Quagmire of Decentralized Machine Learning. Proc. MLSys. (2021)

[7] Li, Q., He, B., Song, D.: Model-Contrastive Federated Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10713-10722 (2021)
[8] Zhu, H., Xu, J., Liu, S., Jin, Y.: Federated Learning on Non-IID Data: A Survey. Neurocomputing. pp. 371-390 (2021)
[9] Mu, M., et al.: FedProc: Prototypical Contrastive Federated Learning on Non-IID Data. Future Generation Computer Systems. pp. 161-173 (2022)
[10] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the International Conference on Machine Learning (ICML). pp. 1597-1607 (2020)
[11] Wu, Y., He, X.: Group Normalization. Proceedings of the European Conference on Computer Vision (ECCV). pp. 3-19 (2018).