NestFL: Efficient Federated Learning through Progressive Model Pruning in Heterogeneous Edge Computing

Xiaomao Zhou
Center of Future Network Research
Purple Mountain Laboratories
Nanjing, China
zhouxiaomao@pmlabs.com.cn

Qingmin Jia
Center of Future Network Research
Purple Mountain Laboratories
Nanjing, China
jiaqingmin@pmlabs.com.cn

Renchao Xie
Center of Future Network Research
Purple Mountain Laboratories
Nanjing, China
xierenchao@pmlabs.com.cn

Abstract

In this paper, we present NestFL, a learning-efficient FL framework for edge computing, which can jointly improve the training efficiency and achieve personalization. Specifically, NestFL takes the runtime resources of the edge devices into consideration and assigns each device a sparse-structured subnetwork by progressively performing the structured pruning. During training, only the updates of these subnetworks are transmitted to the central server. Additionally, these generated subnetworks adopt a structure- and parameter-sharing mechanism, making themselves nested inside a multi-capacity global model. In doing so, the overall communication and computation costs can be significantly reduced, and each device can learn a personalized model without introducing extra parameters. Furthermore, a weighted aggregation mechanism is designed to improve the training performance and maximally preserve personalization.

Keywords: Federated learnings, Model pruning, Multi-capacity model, Personalization

1 Introduction

Federated learning (FL) has been explored as a promising solution for distributed machine learning at the edge. However, the limited capacity and heterogeneity of edge devices usually bring FL with various critical challenges, such as Non-IID data, communication bottleneck, learning inefficiency, etc. In addition, learning a single global model can hardly suffice to work well on all participating devices.

Although many practical applications and theoretical evidences have revealed the potential benefits of FL, several critical challenges for practical FL systems remain unaddressed, such as training DL/ML models on devices with limited resources, ensuring energy efficiency without compromising the learning accuracy, reducing communication burdens, preserving data security and privacy, dealing with data and system heterogeneity, etc. These issues tend to compound when applying FL in edge computing scenarios [1], where devices are usually resource-constrained and heterogeneous.

In addition, most existing FL systems [2] aim to train one single global model for all clients. However, due to the lack of generalization guarantees, the trained global model usually cannot perform well for all devices whose data is statistically heterogeneous. Although producing a personalized model for each participating client is plausible, directly training a global model with multiple sets of parameters will bring in a tremendous amount of computation and communication overheads, which also affects the convergence speed.

In this paper, we propose NestFL, a novel FL framework that can jointly (1) reduce the computation and communication overheads, (2) dynamically adapt to the edge device's available runtime resources, and (3) produce a personalized model for each participating device Specifically, NestFL includes a progressive and distributed model pruning schema into the FL procedure, which produces a heterogeneous and structured-sparse subnetwork for each edge device. During training, NestFL only communicates the updates of these compact subnetworks with the server, thus significantly improving both the computation and communication efficiency. Furthermore, NestFL transforms the global model into a compact multi-capacity model consisting of a set of descendent models, each of which has a unique computation complexity and corresponds to a personalized subnetwork for an edge device. Besides, the multi-capacity model adopts a structure- and parameter-sharing mechanism, which enables the smaller descendent model to share its model parameters with the larger descendent model. As such, the multiple capacities of these descendent models can be nested inside a single global model without introducing extra model parameters. Figure 1 illustrates the architecture of the proposed NestFL framework.

1.1 Progressive Model Pruning

NestFL employs a progressive model pruning strategy to enable each edge device to learn a personalized subnetwork with a sparse structure, where different subnetworks are nested in a single multi-capacity model. Specifically, at the beginning of each communication round, given the global model and the specs of the participating devices, the central server dynamically selects a pruning rate for each device,

ACM MobiCom '22, October 17–21, 2022, Sydney, NSW, Australia

ACM ISBN 978-1-4503-9181-8/22/10.
https://doi.org/10.1145/3495243.3558248

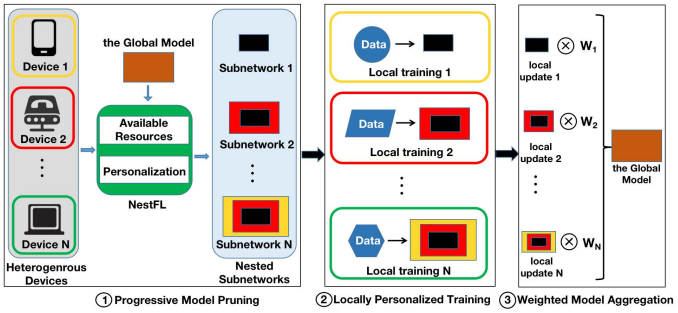


Figure 1. Overview of the proposed NestFL.

and then progressively performs the pruning starting from the global model to generate a set of subnetworks. Starting from the smallest pruning rate, at each pruning step (except for the first step that uses the global model as the initial model), the pruned model from the previous step will be adopted as the initial model to be pruned in the current step. This continues until all the pruned models are generated. In addition, NestFL takes the dynamics of runtime resources into consideration and flexibly modifies the local model to fit the available runtime resources.

1.2 Locally Personalized Training

After receiving the local model, each edge device begins training on it with the local data using Stochastic Gradient Descent (SGD). After performing t steps of SGD, the local update, which embeds the personalized information of the local data distribution, is transmitted to the central server for aggregation. Since only these structure-sparsed and heterogeneous subnetworks are trained and transmitted, the overall computation and communication overheads can be significantly reduced.

1.3 Weighted Model Aggregation

NestFL adopts a novel model aggregation strategy that aims to jointly accelerate the training convergence and preserve personalization under the presence of models with heterogeneous structures. At each aggregation step, only a partition of the global model is aggregated, which starts from the smallest subnetwork, and different partitions are then merged together to update the global model. Additionally, during the aggregation, we measure the correlation between different nodes by calculating the angle between their gradients vectors, based on which a non-linear mapping function is designed to calculate the weights.

2 Early Results

To demonstrate the superiority of NestFL across applications, we experiment with the MNIST [5] and CIFAR10 [6] image classification tasks and the HAR [7] human activity recognition tasks. Following previous works [8], we perform the average shuffling and sorting by class over different datasets to generate the IID data and non-IID data, respectively. We conduct experiments in a simulated setting with 5 clients and a server, where each client is assigned a different computation resource. In addition, each client is assigned IID data or non-IID data with different skewness levels.

We compare our approach with existing federated learning methods with respect to model performance, convergence speed, and the preservation of personalization. The baselines include: (1) FedAvg [3], the classical FL framework that enables distributed learning across multiple devices, where the local models and the global model are designed to share the same architecture. (2) Per-FedAvg [4], a personalized FL approach that incorporates MAML into FedAvg to learn an initial shared model which can be quickly adapted to different clients. (3) HeteroFL [11], a heterogeneous FL solution that coordinatively trains heterogeneous local models with varying computation complexities to produce a single global inference model. (4) Hermes [12], an efficient FL framework that enables each client to learn a personalized and structured-sparse subnetwork.

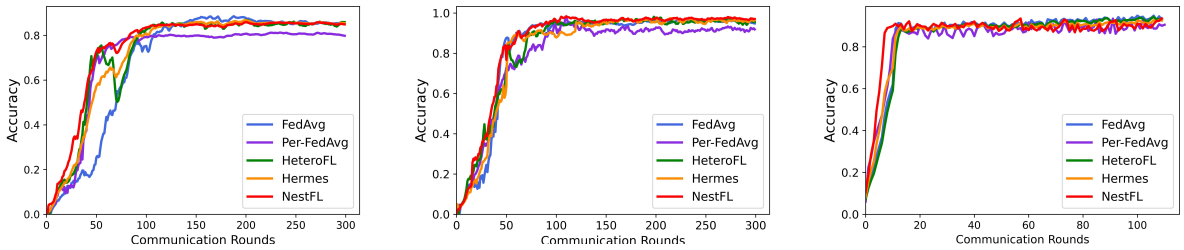
Figure 2 presents the comparison results between NestFL and the baseline methods across different applications. For each application, we plot the convergence of test accuracy across communication rounds. As shown, the proposed NestFL not only consistently achieves better performance than other approaches in all three applications, but also convergence significantly faster (i. e., fewer communication rounds). Additionally, Figure. 3 illustrates the normalized communication cost vs the normalized training time of different approaches. As shown, the proposed NestFL always outperforms other baselines in both the training speed and communication efficiency.

3 Conclusion

In this paper, we propose NestFL, a novel FL framework for edge computing with heterogeneous devices and data, which can significantly reduce the computation and communication overheads and efficiently achieve personalization. Experimental results on two representative FL applications over three datasets demonstrate that NestFL significantly outperforms the status quo approaches in inference accuracy, communication cost, and personalization preservation.

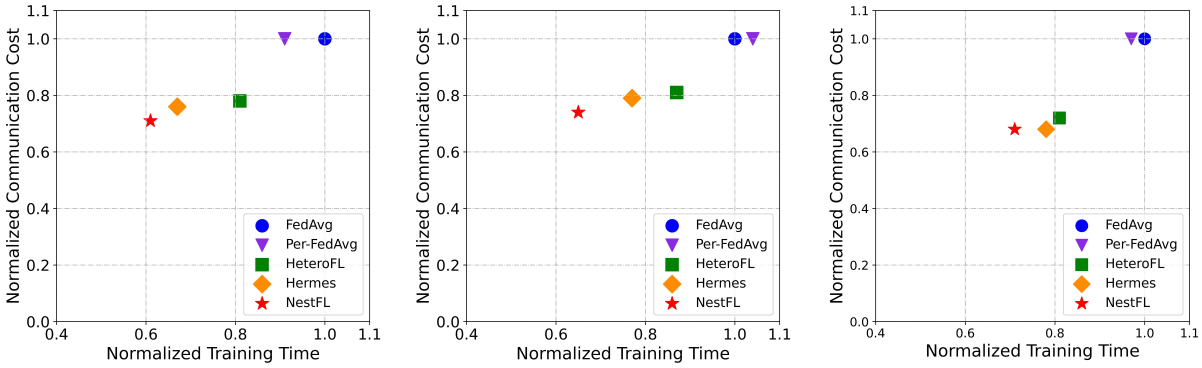
References

[1] Yu, R., Li, P.: Toward resource-efficient federated learning in mobile edge computing. IEEE Network 35(1), pp. 148-155, 2021.
[2] Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10(2), pp. 1-19, 2019.
[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pp. 1273-1282, 2017.
[4] Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning: A meta-learning approach. arXiv preprint arXiv:2002.07948, 2020.



(a) MNIST **(b)** CIFAR10 **(c)** HAR

Figure 2. Performance (Test Accuracy over Communication Rounds) comparison among different approaches.



(a) MNIST **(b)** CIFAR10 **(c)** HAR

Figure 3. Comparison of different approaches in terms of communication costs and training time.

[5] Baldominos, A., Saez, Y., Isasi, P.: A survey of handwritten character recognition with mnist and emnist. Applied Sciences 9(15), p.3169, 2019.

[6] Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images,Technical Report, 2009.

[7] Anguita, D., Ghio, A., Oneto, L., Parra Perez, X. and Reyes Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In Proceedings of the 21th International European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 437-442, 2013.

[8] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[10] Li, A., Sun, J., Wang, B., Duan, L., Li, S., Chen, Y., Li, H.: Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets. arXiv preprint arXiv:2008.03371, 2020.

[11] Diao, E., Ding, J., Tarokh, V.: HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. arXiv preprint arXiv:2010.01264, 2020.

[12] Li, A., Sun, J., Li, P., Pu, Y., Li, H., Chen, Y.: Hermes: an efficient federated learning framework for heterogeneous mobile clients. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, pp. 420-437, 2021.