FedGrav: An Adaptive Federated Aggregation Algorithm for Multi-institutional Medical Image Segmentation

Zhifang Deng[1], Dandan Li[1], Shi Tan[2], Ying Fu[2], Xueguang Yuan[3], Xiaohong Huang[1(✉)], Yong Zhang[4], and Guangwei Zhou[5]

[1] School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, China
huangxh@bupt.edu.cn
[2] Department of Ultrasound, Peking University Third Hospital, Beijing, China
[3] School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[4] Zhongguancun Laboratory, Beijing, China
[5] HTA Co., Ltd., Beijing, China

Abstract. With the increasingly strengthened data privacy acts and the difficult data centralization, Federated Learning (FL) has become an effective solution to collaboratively train the model while preserving each client's privacy. FedAvg is a standard aggregation algorithm that makes the proportion of the dataset size of each client an aggregation weight. However, it can't deal with non-independent and identically distributed (non-IID) data well because of its fixed aggregation weights and the neglect of data distribution. The paper presents a new aggregation strategy called FedGrav, which is designed to handle non-IID datasets and is inspired by the law of universal gravitation in physics. FedGrav can dynamically adjust the aggregation weights based on the training condition of local models throughout the entire training process, making it an effective solution for non-IID data. The model affinity is creatively proposed by considering both the differences of sample size on the client and the discrepancies among local models. It considers the client sample size as the mass of the local model and defines the model graph distance based on neural network topology. By calculating the affinity among local models, FedGrav can explore internal correlations of them and improve the aggregation weights. The proposed FedGrav has been applied to the CIFAR-10 and the MICCAI Federated Tumor Segmentation (FeTS) Challenge 2021 datasets, and the validation results show that our method outperforms the previous state-of-the-art by 1.54 mean DSC and 2.89 mean HD95. The source code will be available on Github.

Keywords: Federated Learning · Brain Tumor Segmentation · FedGrav · Model Affinity · Graph Distance

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 170–180, 2023.
https://doi.org/10.1007/978-3-031-43895-0_16

1 Introduction

The demand for precise medical data analysis has led to the widespread use of deep learning methods in the medical field. However, accompanied by the promulgation of data acts and the strengthening of data privacy, it has become increasingly challenging to train models in large-scale centralized medical datasets. As one of the solutions, federated learning provides a new way out of the dilemma and attracts significant attention from researchers.

Federated learning (FL) [1,2] is a distributed machine learning paradigm in which all clients train a global model collaboratively while preserving their data locally. As a crucial core of them, the aggregation algorithm plays an important role in releasing data potential and improving global model performance. FedAvg [1], as pioneering work, was a simple and effective aggregation algorithm, which makes the proportions of local datasets size as the aggregation weights of local models. But in the real world, not only the numbers of datasets held by clients is different, but also their data distribution may be diverse, which leads to the fact that the data in the federated learning is non-Independent Identically Distribution (non-IID). The naive aggregation algorithms maybe have worse performance because of the non-IID data [3–8]. In medical image segmentation, [9] and [10] took the lead in discussing the application and safety of federated learning in brain tumor segmentation (BraTS). To solve the non-IID challenges of FL in the medical image field, FedDG [11] and FedMRCM [12] were proposed to address the domain shift issue between the source domain and the target domain, but the sharing of latent features may cause privacy concerns. Auto-FedRL [13] and Auto-FedAvg [14] were proposed to deal with the non-IID problem by using an optimization algorithm to learn super parameters and aggregate weights. IDA [15] introduced the Inverse Distance of local models and the average model of all clients to handle non-IID data. The work [16–19] proposed corresponding aggregation methods from the perspectives of clustering, frequency domain, Bayesian, and representation similarity analysis. More than this, the first computational competition on federated learning, Federated Tumor Segmentation (FeTS) Challenge[1] [20] was held to measure the performance of different aggregation algorithms on glioma segmentation [21–24]. Leon et al. [25] proposed FedCostWAvg get a notable improvement compared to FedAvg by including the cost function decreased during the last round and won the challenge. However, most of these methods improve the performance by adding other regular terms to the aggregation method, without considering all factors as a whole, which may limit the performance of the global model.

Different from the above methods, inspired by the concept of the law of universal gravitation in physics, in this paper, we propose a novel aggregation strategy, FedGrav, which unifies the differences in sample size and the discrepancies of local models among clients by defining the concept of model affinity. Specifically, we take the client sample size as the mass of the local model, and the discrepancies among the local models as their distance, which is quantified from

[1] https://fets-ai.github.io/Challenge/.

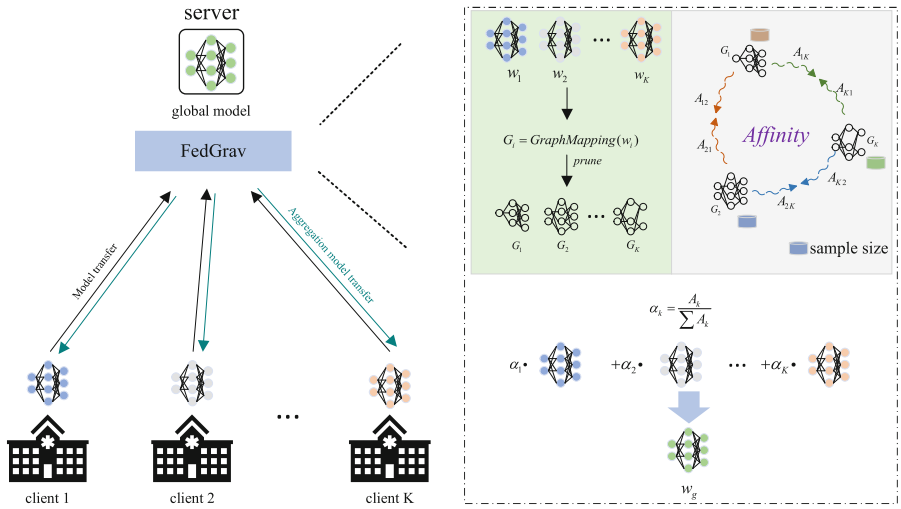

Fig. 1. Overview of the proposed FedGrav. The FedGrav defines the concept of model affinity by unifying the difference in both sample size and local model among clients to aggregates local models and explore the correlations.

the topological perspective of neural networks. Last, the formula 1 is employed to calculate the affinity and explore the internal correlation between the local models. The proposed method promotes a more effective aggregation of local models by unifying the difference between sample size and local model between clients.

The primary contributions of this paper can be summarized as: (1) We propose FedGrav, a novel aggregation strategy that unifies the difference both in sample size and local model among clients by defining the concept of model affinity; (2) We propose Model Graph Distance, a new method to quantify model differences from the perspective of neural network topology. (3) We propose an aggregation algorithm that introduces the concept of affinity and graph into federated learning, and the aggregation weights can be adjusted adaptively; (4) The superior performance is achieved by the proposed method, on the public CIFAR-10 and FeTS challenge datasets.

2 Method

2.1 Overview

Suppose K clients with private data cooperate to train a global model and share the same neural network structure, 3D-Unet [26], which is provided by the FeTS challenge and kept unchanged. For the clients, every client trains a local model w_i for local E epochs and then delivers the local model to the server. The server aggregates local models to a global model by computing the aggregation weights

with the proposed FedGrav and assigns it to all clients. Specifically, given K local models, we first make graph mapping to map the network model to the topology graph, and then the graph distance is obtained after the graph pruning and comparison. For the model affinity computation, FedGrav takes the sample size of every client as the mass of the local model and combines the given graph distance to calculate the affinity between models according to the formula 1. After that, a symmetric Model Affinity Matrix $\mathbb{A} \in \mathbb{R}^{K \times K}$ is analyzed to compute aggregation weights. Last, The server aggregates local models to a global model according to the aggregation weights and assigns it to all clients. Repeat and until T rounds or other limits. An overview of the method is shown in Fig. 1.

2.2 FedGrav

Model Affinity. Inspired by the law of universal gravitation, we assume that there is similar gravitation between any two local models. We define it as model affinity in federated learning. It can be described that the affinity between two local models is proportional to the sample size of the client corresponding to the local model, and inversely proportional to the distance between two models. The equation for model affinity takes the form:

$$A_{ik} = M \frac{n_i n_k}{d_{ik}^2} \tag{1}$$

where A_{ik} is the affinity between i-th and k-th local models, n_i and n_k are the sample size of i-th and k-th client, and d_{ik} is the distance between two local models, which is quantified from the perspective of neural network topology and will be described in the following section. M is the affinity constant, it can be simplified in the subsequent analysis, so this paper will not set specific values for it. The model affinity depicts the internal correlation between two local models, which lays the foundation for accurate aggregation weights.

Graph Distance. The distance is defined to quantify model differences. The differences in local models reflect the discrepancies in the distribution of client data to a certain extent. If the differences in local models can be accurately measured, the more appropriate aggregation weights will be assigned to local models to aggregate a better global model. The key motivation is to measure the internal correlations of local models as accurately as possible. We explore the model distance from the perspective of neural network topology in this paper and define it as model graph distance. In FedGrav, the computation of graph distance goes through the following steps:

(1) Graph Mapping. Suppose the server has received local models trained by local data, and we map them into the topological graph. Inspired by [27], take the j-th convolutional layer of k-th local model with 3D-Unet structure as an example, whose kernel dimension is $3 \times 3 \times 3 \times C_{in} \times C_{out}$, it means this layer has $C_{in} \times C_{out}$ nodes with $3 \times 3 \times 3$ filter, we can obtain $C_{in} \times C_{out}$ weight matrices

of size $3 \times 3 \times 3$. Thus, we get $C_{in} \times C_{out}$ nodes $W \in \mathbb{R}^{27}$. And then, we make every node W as scalar by averaging or summing, which can be formulated as:

$$w_{sum} = \sum_{d=0}^{2} \sum_{h=0}^{2} \sum_{w=0}^{2} W_{dhw}. \tag{2}$$

It can be mapped into a graph whose structure is similar to the full connection layer after the scalarization of the convolutional layer. Given a $3 \times 3 \times 3 \times C_{in} \times C_{out}$ convolutional layer, the dimensions of its input and output are C_{in} and C_{out} respectively. So, we obtain a weight matrix $W_t \in \mathbb{R}^{C_{in} \times C_{out}}$ after averaging or summing the weights of convolution kernel. We take the C_{in} and C_{out} as the number of nodes, and the weight summation w_{sum} is the edge weight.

(2) Graph Pruning. The server collects local models from clients and makes the graph mapping on them to get K graphs which have the same structure except for the edge weights. These graphs contain all the information of local models, including the part of universality and the part of characteristics of the client data. To make the graphs more distinctive, the graph pruning is conducted. In detail, we differentiated these graphs by setting an adaptive threshold δ, where the edge will be removed if the weight difference of each layer between the local models and global model in the last round is less than the threshold, otherwise, the edge will exist. It can be simplified as:

$$edge = \begin{cases} w_{kj}, & \mid w_{kj}^{t} - w_{gj}^{t-1} \mid > \delta, \\ 0, & otherwise. \end{cases} \tag{3}$$

$$\delta = Sort(\mid w_{kj}^{t} - w_{gj}^{t-1} \mid)[\lfloor \lambda \cdot C_{in} \times C_{out} \rfloor], 0 \leq \lambda < 1. \tag{4}$$

where in Eq. 3, 0 denotes the edge is removed, w_{kj}^{t} denotes edge weight of the j-th layer from the k-th graph in the t-th round, also the weight summation of the j-th layer from the k-th local model in the t-th round, w_{gj}^{t-1} is the weight summation of the j-th layer from the global model in $(t-1)$-th round. The threshold δ varies adaptively with the weights of local models, and λ is the pruning ratio which is responsible for adjusting the degree of pruning. After that we get K discriminative graphs G_i, $i \in [1, K]$.

(3) Graph Comparison. In order to measure the degree of correlation between two graphs, we measure the similarity between pairs of graphs by computing matching between their sets of embeddings, where the Pyramid Match Graph Kernel [28] is employed. We take the reciprocal of the correlation degree as the distance between them. The distance is defined as follows:

$$d_{ik} = \frac{1}{PyramidMatch(G_i, G_k)} \tag{5}$$

Aggregation Weights. According to the above process, the Affinity Matrix $\mathbb{A}$ is obtained, which reports the correlation among local models and is symmetric. The element A_{ik} in matrix $\mathbb{A}$ denotes the affinity of G_i and G_k. The elements in

Table 1. Comparisons with other state-of-the-art methods on the CIFAR-10 dataset.

Method	Accuracy (%)
FedAvg [1]	88.37 ± 0.04
FedProx [6]	87.93 ± 0.19
FedNova [29]	88.68 ± 0.26
Auto-FedAvg [14]	89.16
FedGrav	$\mathbf{89.35 \pm 0.23}$

the k-th row represent the Affinity among G_k and all graphs, so we can get the affinity of the k-th graph with all graphs, which denotes the correlations of G_k with the whole graphs. last, we normalize A_k as the aggregation weight of the k-th local model or layer.

$$\alpha_k = \frac{\sum_{i=1}^{K} A_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{K} A_{ik}} \tag{6}$$

In federated learning, clients send the updated local models back to the server each round. In round t, α_k is represented as α_k^t. The global model w_g^{t+1} is aggregated by the server:

$$w_g^t = \sum_{k=1}^{K} \alpha_k^t \cdot w_k^t \tag{7}$$

then, the server assigns the global model w_g^t to all clients. Repeat and until T rounds or other limits.

3 Experiments

3.1 Datesets and Settings

CIFAR-10. The first dataset to verify the validity of our algorithm is CIFAR-10. We partition the training set into 8 clients with heterogeneous data by sampling from a Dirichlet distribution ($\alpha = 0.5$) as in [10] to simulate the non-IID distribution, and the test set in CIFAR-10 is considered as the global test set to evaluate the performance of different algorithms. VGG-9 [30] is employed for image classification, and the other detailed settings are as follows: initial learning rate of $1e-2$; total rounds of 100; local epochs of 20; batch size of 64; SGD optimizer for clients.

MICCAI FeTS2021 Training Data. The real-world dataset used in experiments is provided by the FeTS Challenge organizer, which is the training set of the whole dataset about brain tumor segmentation. In order to evaluate the performance of FedGrav, we partition the dataset composed of 341 data samples

Table 2. Comparisons with other state-of-the-art methods on the MICCAI FeTS2021 Training dataset. D denotes DICE, H95 denotes HD95, and M denotes mean.

Method	D WT	D ET	D TC	H95 WT	H95 ET	H95 TC	M D	M H95
FedAvg [1]	90.49	73.03	69.38	4.82	33.88	39.00	77.63 ± 0.573	25.90 ± 2.731
FCW [21]	90.88	73.15	70.56	3.74	40.79	**17.16**	78.20 ± 0.749	20.56 ± 0.311
FedGrav	**91.26**	**77.21**	**70.75**	**2.80**	**27.40**	22.8	$\mathbf{79.74 \pm 0.595}$	$\mathbf{17.67 \pm 1.692}$

into training set and validation set according to the ratio of 8 : 2, and the data is unevenly distributed between 17 data clients. The segmentation network, 3D-Unet, is provided by FeTS and kept unchanged, the learning rate is $1e-4$ and the local epochs are 10. Limited by the framework and official code mechanism, the total number of rounds of training is set to 70, although the performance of the algorithm does not converge to the best.

3.2 Results

Experiment Results on the CIFAR-10. We first validate the proposed method on the CIFAR-10 dataset. Table 1 shows the quantitative results of the state-of-the-art FL methods in terms of the average accuracy, such as FedAvg [1], FedProx [6], FedNova [29], and Auto-FedAvg [14]. As can be seen from the table, the proposed FedGrav method outperforms the other competing FL aggregation methods including Auto-FedAvg, a learning-based aggregation method, which indicates the potential and superiority of FedGrav.

Experiment Results on MICCAI FeTS2021 Training Dataset. In order to verify the robustness of our method and its performance in real-world data, we conduct the experiment on the MICCAI FeTS2021 Training dataset. We evaluate the performance of our algorithm by comparing six indicators: the Dice Similarity Coefficient(DSC) and Hausdorff Distance-95th percentile(HD95) of whole tumor(WT), enhancing tumor(ET), and tumor core(TC). As is shown in Table 2, we list the average results of FedAvg, FedCostWAvg(shortened to FCW), the champion method of FeTS Challenge 2021, and the proposed FedGrav. Different from the original FedCostWAvg which changed the activation function of networks, our re-implemented version made the network unchanged to ensure a fair comparison. Through the quantitative comparison in Table 2, we can find that the proposed method FedGrav has achieved the best results in all indicators except the HD95 TC. Moreover, compared with FedCostWAvg, FedGrav has significantly improved the evaluation of segmentation performance, especially in the enhancing tumor segmentation.

The visualization results are shown in Fig. 2. It can be seen that our FedGrav achieves better segmentation results, even in the hard example, compared to FedCostWAvg and FedAvg. The results proved that the proposed method FedGrav can explore the correlations of local models better and achieved more excellent aggregation performance compared with other methods.

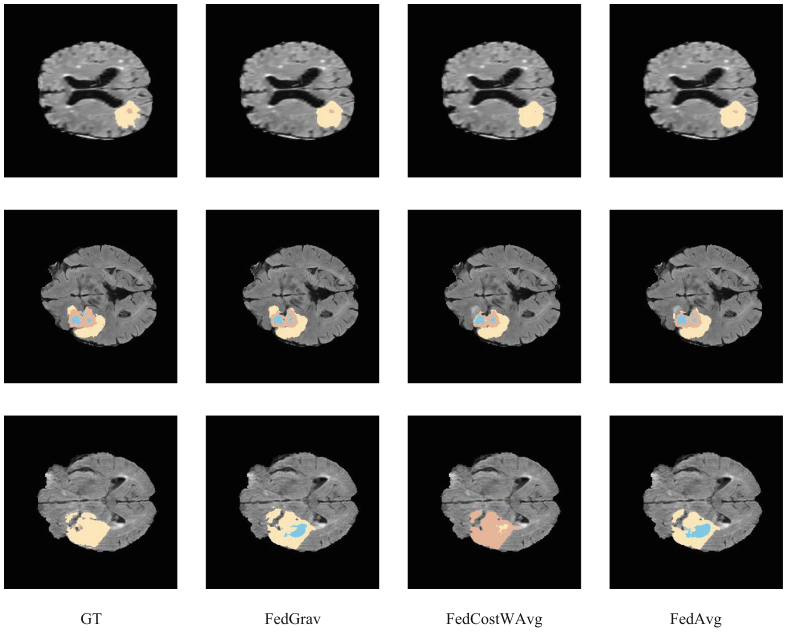

Fig. 2. The visual comparisons with previous state-of-the-art methods on the MICCAI FeTS2021 Training dataset.

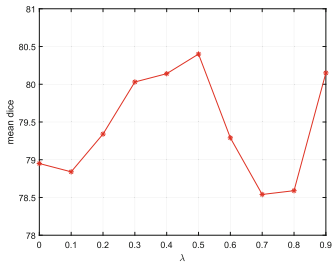

Fig. 3. Comparison of different pruning ratio λ in FedGrav on FeTS datasets.

3.3 Ablation Study

To evaluate the effectiveness and find the better configuration of FedGrav, we conduct the ablation study on the FeTS datasets, and the results are shown in Fig. 3. As we can see, the mean DSC shows a trend of rising first and then falling, because more irrelevant and redundant information will be saved in the model when pruning is not performed. The different values of λ denote the loose degree of graphs, with the gradual increase of λ, the redundant information in local models is gradually eliminated, and the unique information of each local model is preserved. While, when the pruning ratio λ increases to a certain extent, the

models lack key information, which makes the model affinity inaccurate, resulting in a decline in segmentation performance.

4 Conclusion

In this paper, we introduced FedGrav, a novel aggregation strategy inspired by the law of universal gravitation in physics. FedGrav improves local model aggregation by considering both the differences in sample size and discrepancies among local models. It can adaptively adjust the aggregation weights and explore the internal correlations of local models more effectively. We evaluated our method on CIFAR-10 and real-world MICCAI Federated Tumor Segmentation Challenge (FeTS) datasets, and the superior results demonstrated the effectiveness and robustness of our FedGrav.

Acknowledgements. This work was supported by the Fund for Innovation and Transformation of Haidian District, Beijing, China(No. HDCXZHKC2021201)

References

1. 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. PMLR (2017)
2. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 1–19 (2019)
3. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-IID data. arXiv preprint arXiv:1806.00582 (2018)
4. Li, X., Jiang, M., Zhang, X., et al.: FedBN: federated learning on non-IID features via local batch normalization. In: International Conference on Learning Representations (2020)
5. Li, T., Sahu, A.K., Zaheer, M., et al.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. **2**, 429–450 (2020)
6. Sattler, F., Wiedemann, S., Maluller, K.-R., Samek, W.: Robust and communication-efficient federated learning from non-IID data. IEEE Trans. Neural Networks Learn. Syst. **31**, 3400–3413 (2019)
7. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.: Scaffold: stochastic controlled averaging for federated learning. ICML 2020 (2020)
8. Chen, X., Chen, T., Sun, H., Wu, Z.S., Hong, M.: Distributed training with heterogeneous data: bridging median- and mean-based algorithms. In: NeurIPS 2020 (2020)
9. Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S.: Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 92–104. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_9
10. Li, W., et al.: Privacy-preserving federated brain tumour segmentation. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 133–141. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_16

11. Liu, Q., Chen, C., Qin, J., et al.: Feddg: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1013–1023 (2021)
12. Guo, P., Wang, P., Zhou, J., Jiang, S., Patel, V.M.: Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2423–2432 (2021)
13. Guo, P., et al.: Auto-FedRL: federated hyperparameter optimization for multi-institutional medical image segmentation. arXiv preprint arXiv:2203.06338 (2022)
14. Xia, Y., Yang, D., Li, W., et al.: Auto-FedAvg: learnable federated averaging for multi-institutional medical image segmentation. arXiv preprint arXiv:2104.10195 (2021)
15. Yeganeh, Y., Farshad, A., Navab, N., Albarqouni, S.: Inverse distance aggregation for federated learning with non-IID data. In: Albarqouni, S., et al. (eds.) DART/DCL -2020. LNCS, vol. 12444, pp. 150–159. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60548-3_15
16. Palihawadana, C., Wiratunga, N., Wijekoon, A., et al.: FedSim: similarity guided model aggregation for Federated Learning. Neurocomputing **483**, 432–445 (2022)
17. Chen, H.Y., Chao, W.L.: FedBE: making Bayesian model ensemble applicable to federated learning. In: International Conference on Learning Representations
18. Chen, Z., Zhu, M., Yang, C., Yuan, Y.: Personalized retrogress-resilient framework for real-world medical federated learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 347–356. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_33
19. Dong, N., Voiculescu, I.: Federated contrastive learning for decentralized unlabeled medical images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 378–387. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_36
20. Pati, S., et al.: The federated tumor segmentation (fets) challenge. arXiv preprint arXiv:2105.05874 (2021)
21. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**(1), 1–13 (2017)
22. Reina, G.A., et al.: Open: an open-source framework for federated learning. arXiv preprint arXiv:2105.06413 (2021)
23. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**(1), 1–12 (2020)
24. Koer, F., et al.: Brats toolkit: translating brats brain tumor segmentation algorithms into clinical and scientific practice. Front. Neurosci. **14**, 125 (2020)
25. Mächler, L., Ezhov, I., Kofler, F., et al.: FedCostWAvg: a new averaging for better Federated Learning. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021, Part II. LNCS, vol. 12963, pp. 383–391. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-09002-8_34
26. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
27. Gabrielsson, R.B.: Topological Data Analysis of Convolutional Neural Networks' Weights on Images

28. Nikolentzos, G., Meladianos, P., Vazirgiannis, M.: Matching node embeddings for graph similarity. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, pp. 2429–2435 (2017)
29. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)