FedSoup: Improving Generalization and Personalization in Federated Learning via Selective Model Interpolation

Minghui Chen[1], Meirui Jiang[2], Qi Dou[2], Zehua Wang[1], and Xiaoxiao Li[1(✉)]

[1] Department of Electrical and Computer Engineering,
The University of British Columbia, Vancouver, Canada
`xiaoxiao.li@ece.ubc.ca`
[2] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China

Abstract. Cross-silo federated learning (FL) enables the development of machine learning models on datasets distributed across data centers such as hospitals and clinical research laboratories. However, recent research has found that current FL algorithms face a trade-off between local and global performance when confronted with distribution shifts. Specifically, personalized FL methods have a tendency to overfit to local data, leading to a sharp valley in the local model and inhibiting its ability to generalize to out-of-distribution data. In this paper, we propose a novel federated model soup method (*i.e.*, selective interpolation of model parameters) to optimize the trade-off between local and global performance. Specifically, during the federated training phase, each client maintains its own global model pool by monitoring the performance of the interpolated model between the local and global models. This allows us to alleviate overfitting and seek flat minima, which can significantly improve the model's generalization performance. We evaluate our method on retinal and pathological image classification tasks, and our proposed method achieves significant improvements for out-of-distribution generalization. Our code is available at https://github.com/ubc-tea/FedSoup.

1 Introduction

Federated learning (FL) has emerged as a promising methodology for harnessing the power of private medical data without necessitating centralized data governance [6,7,22,25]. However, recent study [28] has identified a significant issue in current FL algorithms, namely, the trade-off between local and global performance when encountering distribution shifts. This issue is particularly

prevalent in medical scenarios [12,16,17], where medical images may undergo shifts of varying degrees due to differences in imaging device vendors, parameter settings, and the patient demographics. Personalized FL (PFL) techniques are typically utilized to address the data heterogeneity problem by weighing more on in-distribution (ID) data of each client. For instance, FedRep [5] learns the entire network during local updates and keeps part of the network from global synchronization. However, they have a risk of overfitting to local data [23], especially when client local data is limited, and have poor generalizability on out-of-distribution (OOD) data. Another line of work has studied the heterogeneity issue by regularizing the updates of local model. For instance, FedProx [15] constraints local updates to be closer to the global model. An effective way to evaluate FL's generalizability is to investigate its performance on the joint global distribution following [28], which refers to testing the FL models on $\bigcup\{\mathcal{D}_i\}$, where $\mathcal{D}_i$ indicates client i's distribution[1]. Unfortunately, the existing works have not found the sweet spot between personalized (local) and consensus (global) models.

In this regard, we study a practical problem of enhancing personalization and generalization jointly in cross-silo FL for medical image classification when faced data heterogeneity. To this end, we aim to address the following two questions in FL: *What could be the causes that result in local and global trade-off?* and *How to achieve better local and global trade-off?* First, we provide a new angle to understand the trade-off. We reveal that over-personalization in FL can cause overfitting on local data and trap the model into a sharp valley of loss landscape (highly sensitive to parameter perturbation, see detailed definition in Sec. 2.2), thus limiting its generalizability. An effective strategy for avoiding sharp valleys in the loss landscape is to enforce models to obtain flat minima. In the centralized domain, weight interpolation has been explored as a means of seeking flat minima as its solution is moved closer to the centroid of the high-performing models, which corresponds to a flatter minimum [3,6,11,24]. However, research on these interpolation methods has been overlooked in FL.

With the above basis, we propose to track both local and global models during the federated training and perform model interpolation to seek the optimal balance. Our insight is drawn from the model soup method [27], which shows that averaging weights of multiple trained models with same initial parameters can enhance model generalization. However, the original model soup method requires training substantial models with varying hyper-parameters, which can be prohibitively time-consuming and costly in terms of communication during FL. Given the high communication cost and the inability to restart training from scratch in FL, we leverage global models at different time points within a single training session as the ingredients to adapt the model soup method [27] to FL.

In this paper, we propose a novel federated model soup method (FedSoup) to produce an ensembled model from local and global models that achieve better local-global trade-off. We refer the 'soup' as a combo of different federated models. Our proposed FedSoup includes two key modules. The first one is temporal model selection, which aims to select suitable models to be combined into

---

[1] In the heterogeneous setting ($\mathcal{D}_i \neq \mathcal{D}_j$), $\mathcal{D}_j$ is viewed as the OOD data for client i.
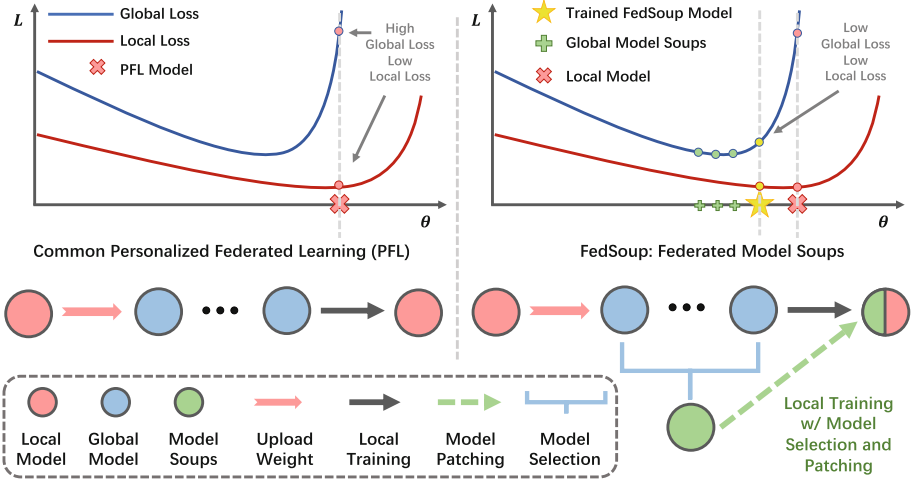
Fig. 1. The overview of our method (FedSoup) compared with common PFL methods. PFL methods typically minimize local loss but suffering high global loss. While our federated model soup method balances both local and global loss by seeking flat minima. The black dots in the figure represent ellipsis and indicate multiple rounds of model upload and model training in between. Compared to previous pFL methods, our approach introduces global model selection modules and local model interpolation with the global model (referred to as model patching).

one. The second module is Federated model patching [10], which refers to a fine-tuning technique that aims to enhance personalization without compromising the already satisfactory global performance. For the first module, temporal model selection, we utilize a greedy model selection strategy based on the local validation performance. This avoids incorporating models that could be located in a different error landscape basin than the local loss landscape (shown in Fig. 1). Consequently, each client possesses their personalized global model soups, consisting of a subset of historical global models that are selected based on their local validation sets. As for the second module, federated model patching, it introduces model patching in local client training by interpolating the local model and the global model soups into a new local model, bridging the gap between local and global domains. It promotes the personalization of the model for ID testing and also maintains good global performance for OOD generalization.

In summary, our key contributions are as follows: (i) A novel FL method called Federated Model Soups (FedSoup) is proposed to improve generalization and personalization by promoting smoothness and seeking flat minima. (ii) A new temporal model selection mechanism is designed for FL, which maintains a client-specific model soups with temporal history global model to meet personalization requirements while not incurring additional training costs. (iii) An innovative federated model patching method between local and global models is introduced in federated client training to alleviate overfitting of local limited data.

2   Method

2.1   Problem Setup

Consider a cross-silo FL setting with N clients. Let $\mathcal{D} := \{\mathcal{D}_i\}_{i=1}^N$ be a set of N training domain, each of which is a distribution over the input space $\mathcal{X}$. For each client, we have access to n training data points in the form of $(x_j^i, y_j^i)_{j=1}^n \sim \mathcal{D}_i$, where y_j^i denotes the target label for input x_j^i. We also define a set of unseen target domains $\mathcal{T} := \{\mathcal{T}_i\}_i^{N'}$ in a similar manner, where N' is the number of target domains and is typically set to one. The goal of personalization (local performance) is to find a model $f(\cdot; \theta)$ via minimizing an empirical risk $\widehat{\mathcal{E}}_{\mathcal{D}_i}(\theta) := \frac{1}{n} \sum_{j=1}^n \ell(f(x^i; \theta), y^i))$ over a local client training set $\mathcal{D}_i$, where $\ell(\cdot, \cdot)$ is a loss function. On the other hands, the objective of generalization (global performance) is to minimize both population loss $\mathcal{E}_{\mathcal{D}}(\theta)$ and $\mathcal{E}_{\mathcal{T}}(\theta)$ over multiple domains by Empirical Risk Minimization (ERM) $\widehat{\mathcal{E}}_{\mathcal{D}}(\theta) := \frac{1}{Nn} \sum_{i=1}^N \sum_{j=1}^n \ell(f(x^i; \theta), y^i))$ over all training clients' training sets $\mathcal{D}$. In this work, we evaluate the local performance on local testing samples from $\mathcal{D}_i$, and evaluate the global performance on testing samples from the joint global distribution $\mathcal{D} := \{\mathcal{D}_i\}_{i=1}^N$.

2.2   Generalization and Flat Minima

In practice, ERM in deep neural networks, *i.e.*, $\arg\min_\theta \widehat{\mathcal{E}}_{\mathcal{D}}(\theta)$, can yield multiple solutions that offer comparable training loss, but vastly different levels of generalizability [3]. However, without proper regularization, models are prone to overfit the training data and the training model will fall into a sharp valley of the loss surface, which is less generalizable [4].

One common reason for failures in ERM is the presence of variations in the data distribution ($\mathcal{D}_i \neq \mathcal{D}$), which can cause a shift in the loss landscape. As illustrated in Fig. 1, the sharper the optimized minima, the more sensitive it is to shifts in the loss landscape. This results in an increase in generalization error. In cross-silo FL, each client may overfit their local training data, leading to poor global performance. This is due to the distribution shift problem, which creates conflicting objectives among the local models [23]. Therefore, when the local model converges to a sharp minima, the higher the degree of personalization (local performance) of the model, the more likely it is to have poor generalization ability (global performance).

From the domain generalization formalization in [2,3], the test loss $\mathcal{E}_{\mathcal{T}}(\theta)$ can be bounded by the robust empirical loss $\widehat{\mathcal{E}}_{\mathcal{D}}^\epsilon(\theta)$ as follows:

$$\mathcal{E}_{\mathcal{T}}(\theta) < \widehat{\mathcal{E}}_{\mathcal{D}}^\epsilon(\theta) + \frac{1}{N} \sum_{i=1}^N \sup_{A \in \mathcal{A}} |\mathcal{P}_{\mathcal{D}_i}(A) - \mathcal{P}_{\mathcal{T}}(A)| + \xi, \tag{1}$$

where the $\sup_A |\mathcal{P}_{\mathcal{D}_i}(A) - \mathcal{P}_{\mathcal{T}}(A)|$ is a divergence between domain $\mathcal{D}_i$ and $\mathcal{T}$, $\mathcal{A}$ is the set of measurable subsets under $\mathcal{D}_i$ and $\mathcal{T}$, and ξ is the confidence bound term related to the radius ϵ and the number of the training samples.

From the Equation (1), we can infer that minimizing sharpness and seeking flat minima is directly related with the generalization performance on the unseen target domain.

2.3   Our Solution: FedSoup

After analyzing the aforementioned relationship between sharpness and generalization, we expound on the distinctive challenges of seeking flat minima and mitigating the trade-off between local and global performance in FL. Consequently, we introduce two refined modules as the ingredient of our proposed FedSoup solution. FedSoup only needs to modify the training method of the client, and the algorithm implementation is shown in Algorithm 1.

Temporal Model Selection. Stochastic Weight Averaging (SWA) [11] and Sharpness-Aware Minimization (SAM) [8] are two commonly used flat minima optimizers, which seek to find parameters in wide low-loss basins. In contrast to SAM, which incurs extra computational cost to identify the worst parameter perturbation, SWA is a more succinct and effective approach for implicitly favoring the flat minima by averaging weights. The SWA algorithm is motivated by the observation that SGD often finds high-performing models in the weight space but rarely reaches the central points of the optimal set. By averaging the parameter values over iterations, the SWA algorithm moves the solution closer to the centroid of this space of points.

Nevertheless, when it comes to cross-silo FL training, the discrepancy between clients is significant, and models might lie in different basins. Merging all these models haphazardly is not effective and might hinder generalization. Recently, a selective weight averaging method called model soups [27] was introduced to enhance the generalization of fine-tuned models. The original model soups is not applicable to the FL setting, requiring high communication costs and training compute. We adapt the idea to a new approach by leveraging global models trained at different time points in one pass of FL training. Additionally, considering the heterogeneity of data distribution in cross-silo FL and the requirement of personalization, we propose a model selection strategy where each client utilizes the performance of its local validation set as a monitoring indicator. We called this module temporal model selection (see Algorithm 1 Line 7-8)

Federated Model Patching. According to previous analysis on the loss landscape, there is a loss landscape offset between different FL clients due to their domain discrepancy. Thus, simply integrating a global model can damage the model's personalization. To address this, we introduce the use of the model patching [10] (*i.e.* local and global model interpolation) during client-side local training in FL, aiming to improve model personalization and maintain the good global performance. Specifically, model patching approach forces local client not to distort global model severely and seek low-loss interpolated model between local and global, encouraging the local and global model lie in the same basin

without a large barrier of linear connectivity. [19]. We called this module federated model patching. In summary, the update rule of FedSoup is implemented as follows:

$$\theta_{FedSoup} \leftarrow \frac{\theta_g^1 + \cdots + \theta_g^k + \theta_l}{k+1}, \tag{2}$$

where θ_g is global model, θ_l is local model, k is the number of selected global models. This update rule corresponds to Algorithm 1 Line 9.

Algorithm 1. FedSoup

1: **Input:** global model θ_g, local model θ_l, last epoch local model $\widehat{\theta_l^i}$, number of clients k, number of epochs n, interpolation start epoch E.
2: soup $\leftarrow \{\}$
3: **for** $i = 0$ **to** n **do**
4:     $\theta_g \leftarrow$ Aggregation($\widehat{\theta_l^1}, \ldots, \widehat{\theta_l^k}$)
5:     $\theta_l \leftarrow$ ClientUpdate(θ_g)
6:     **if** $i \geq E$ **then**
7:         **if** ValAcc(average(soup $\cup \{\theta_l\} \cup \{\theta_g\}$)) $\geq$ ValAcc(average(soup $\cup \{\theta_l\}$)) **then**
8:             soup $\leftarrow \{\theta_g\}$ {Module 1: Temporal Model Selection}
9:         $\theta_l \leftarrow$ average(soup $\cup \{\theta_l\}$)   {Module 2: Federated Model Patching}
10: **return** θ_l

It is important to note that our proposed FedSoup algorithm requires only one carefully tuned hyper-parameter, namely the interpolation start epoch. To mitigate the risk of having empty global model soups when the start epoch is too late and to prevent potential performance degradation when the start epoch is too early, we have set the default interpolation start epoch to be 75% of the total training epochs, aligning with the default setting of SWA. Furthermore, it is worth mentioning that the modified model soup and model patching modules in our proposed FedSoup framework are interdependent. Model patching, which is a technique based on our modified model soup algorithm, provides an abundance of models to explore flatter minima and enhance performance.

3 Experiments

3.1 Experimental Setup

Dataset. We validate the effectiveness of our proposed method, FedSoup, on two medical image classification tasks. The first task involved the classification of pathology images from five different sources using Camelyon17 dataset [1], and each source is viewed as a client. The pathology experiment consists of a total of $4,600$ images[2], each with a resolution of 96×96. The second task

---

[2] We take a random subset from the original Camelyon17 dataset to match the small data settings in FL [18].

involved retinal fundus images from four different institutions [9,21,26], and each institute is viewed as a client. The retinal fundus experiment consists of a total of 1,264 images, each with a resolution of 128×128. The objective of both datasets is to identify abnormal images from normal ones. We also maintained an equal number of samples from each client to prevent clients with more data from having a disproportionate influence on the global performance evaluation.

Table 1. Local and global performance results comparison with SOTA PFL methods on two medical image classification tasks.

Method	Pathology				Retinal Fundus			
	Local Performance		Global Performance		Local Performance		Global Performance	
	Accuracy ↑	AUC ↑	Accuracy ↑	AUC ↑	Accuracy ↑	AUC ↑	Accuracy ↑	AUC ↑
FedAvg [18]	$82.41_{(0.61)}$	$90.44_{(0.42)}$	$70.18_{(1.25)}$	$78.74_{(1.31)}$	$89.81_{(1.17)}$	$95.24_{(0.73)}$	$73.27_{(3.97)}$	$81.02_{(4.57)}$
FedProx [15]	$\mathbf{86.34}_{(0.52)}$	$\mathbf{92.78}_{(0.36)}$	$67.42_{(1.32)}$	$77.18_{(1.32)}$	$90.14_{(0.20)}$	$95.47_{(0.63)}$	$70.46_{(2.36)}$	$76.84_{(2.93)}$
MOON [14]	$85.71_{(0.42)}$	$91.98_{(0.34)}$	$70.61_{(1.54)}$	$79.27_{(1.43)}$	$89.14_{(0.77)}$	$94.91_{(0.98)}$	$77.49_{(3.15)}$	$84.89_{(2.70)}$
FedBN [17]	$82.32_{(0.87)}$	$90.07_{(0.77)}$	$65.16_{(0.69)}$	$71.61_{(0.97)}$	$89.70_{(1.45)}$	$95.49_{(0.56)}$	$75.11_{(0.52)}$	$82.70_{(1.01)}$
FedFomo [30]	$80.99_{(1.07)}$	$86.51_{(0.99)}$	$61.00_{(0.59)}$	$61.69_{(0.78)}$	$89.70_{(0.00)}$	$94.88_{(0.45)}$	$59.90_{(2.00)}$	$63.82_{(1.14)}$
FedRep [5]	$82.50_{(0.53)}$	$89.77_{(0.43)}$	$66.87_{(0.60)}$	$72.34_{(0.75)}$	$89.38_{(0.89)}$	$95.16_{(0.83)}$	$71.81_{(1.79)}$	$79.09_{(2.44)}$
FedBABU [20]	$85.18_{(0.33)}$	$92.39_{(0.29)}$	$69.56_{(1.40)}$	$77.26_{(1.48)}$	$90.25_{(1.39)}$	$95.10_{(0.33)}$	$77.65_{(0.24)}$	$85.09_{(0.21)}$
FedSoup	$85.71_{(0.37)}$	$92.47_{(0.31)}$	$\mathbf{72.87}_{(1.35)}$	$\mathbf{81.45}_{(1.40)}$	$\mathbf{90.92}_{(0.50)}$	$\mathbf{96.00}_{(0.43)}$	$\mathbf{78.64}_{(0.90)}$	$\mathbf{86.24}_{(0.86)}$

Evaluation Setup. For each client, we take 75% of the data as the training set. To assess the generalization ability and personalization of our model, we have constructed both local and global testing sets. Following [28], in our experimental setting, we first create a held-out global testing set by randomly sampling an equal number of images per source/institute, so its distribution is different from either of the client. The local testing dataset for each FL client is the remaining sample from the same source as its training set. The number of local testing set per client is approximately the same as that of the held-out global testing set. For the pathology dataset, as each subject can have multiple samples, we have ensured that data from the same subject only appeared in either the training or testing set. To align with the cross-validation setting for subsequent out-of-domain evaluations, we conducted a five-fold leave-one-client-data cross-validation, with three repetitions using different random seeds in each fold. The results of the repeated experiments without hold-out client data are provided in the appendix. For PFL methods, we report the performance by averaging the results of each personalized models.

Models and Training Hyper-Paramters. We employ the ResNet-18 architecture as the backbone model. Our approach initiates local-global interpolation at the 75% training phase, consistent with the default hyper-parameter setting of SWA. We utilize the Adam optimizer with learning rate of $1e-3$, momentum coefficients of 0.9 and 0.99 and set the batch size to 16. We set the local training epoch to 1 and perform a total of 1,000 communication rounds.

3.2   Comparison with State-of-the-Art Methods

We compare our method with seven common FL and state-of-the-art PFL methods. Results in Table 1 demonstrate that our FedSoup method achieves competitive performance on local evaluation sets while significantly improving generalization with respect to global performance. Furthermore, FedSoup exhibits greater stability with lower variance of performance across multiple experiments.
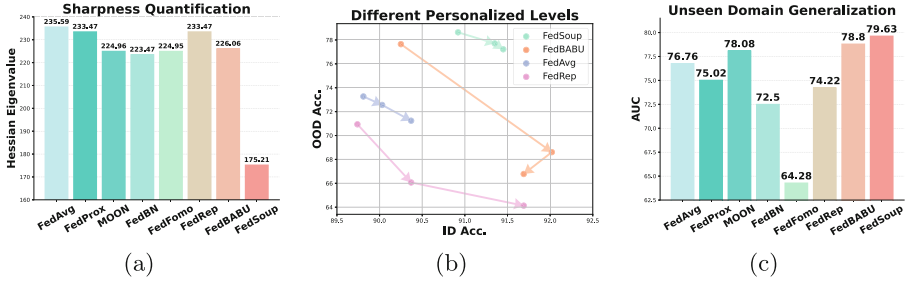

Fig. 2. Analysis of our approach: (a) sharpness quantification on the retina dataset, (b) local and global trade-off under different personalized levels (fine-tuning epochs), (c) unseen domain generalization on the pathology dataset.

Comparing the performance improvement of FedSoup across different datasets, we observed that FedSoup had a more substantial effect on the smaller retinal fundus dataset compared to the larger pathology dataset. In terms of performance gap compared to the second-best methods (FedBABU on Retina and FedProx on Pathology), our approach demonstrates a larger advantage on the Retina dataset. This observation suggests that our proposed method can mitigate the negative impact of local overfitting caused by small local datasets, thus improving the model's generalization ability.

Sharpness Quantification. The sharpness measure used in this study calculates the median of the dominating Hessian eigenvalue across all training set batches using the Power Iteration algorithm [29]. This metric indicates the maximum curvature of the loss landscape, which is often used in the literature on flat minima [13] to reflect the sharpness. The median of the dominating Hessian eigenvalue of all clients in the retinal fundus dataset was measured in this part. Based on the Fig. 2(a) presented, it is evident that the proposed method leads to flatter minima as compared to the other methods.

Trade-off at Different Personalized Levels. Following the evaluation in [28], we conducted an experiment comparing the local and globalperformance of different models at various levels of personalization using the retinal fundus datasets. We control the personalization level for the PFL methods by varying the number of iterations that the model undergoes fine-tuning using only the

local training set after federated training. As we increase the number of fine-tuning iterations, we consider the level of personalization to be higher. We choose to show model performance after fine-tuning iterations 1, 7, and 15. The results in Fig. 2(b) indicated that existing FL methods often have a trade-off between local and global performance. As the number of fine-tuning iterations increases, local performance typically improves but global performance decreases. Compared to other methods, our approach maintains high local performance while also preventing a significant drop in global performance, which remains much higher than other methods.

3.3   Unseen Domain Generalization

We show the additional benefits of FedSoup on unseen domain generalization.

Setup. To evaluate the generalization of our method beyond the participating domains, we utilize one domain that did not take part in the distributed training and used its data as the evaluation set for unseen domain generalization. To this end, we perform leave-one-out cross-validation by having one client as the to-be-evaluated unseen set each time. To ensure a reliable results of unseen domain generalization, we conducted experiments on the Camelyon17 dataset, which has a larger number of samples.

Results. Overall, our proposed method demonstrates an advantage in terms of unseen domain generalization capabilities (see Fig. 2(c)). In comparison to FedAvg, Our approach resulted in a 2.87-point increase in the AUC index for generalization to unseen domains on the pathology dataset.

4   Conclusion

In this paper, we demonstrate the trade-off between personalization and generalization in the current FL methods for medical image classification. To optimize this trade-off and achieve flat minima, we propose the novel FedSoup method. By maintaining personalized global model pools in each client and interpolating weights between local and global models, our proposed method enhances both generalization and personalization. FedSoup outperforms other PFL methods in terms of both generalization and personalization, without incurring any additional inference or memory costs.

References

1. Bándi, P., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans. Med. Imaging **38**(2), 550–560 (2019)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**(1–2), 151–175 (2010)

3. Cha, J., et al.: SWAD: domain generalization by seeking flat minima. In: NeurIPS, pp. 22405–22418 (2021)
4. Chaudhari, P., et al.: Entropy-SGD: biasing gradient descent into wide valleys. In: ICLR (Poster). OpenReview.net (2017)
5. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 2089–2099. PMLR (2021)
6. Dayan, I., et al.: Federated learning for predicting clinical outcomes in patients with covid-19. Nat. Med. **27**(10), 1735–1743 (2021)
7. Dou, Q., So, T.Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., et al.: Federated deep learning for detecting covid-19 lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ Digit. Med. **4**(1), 1–11 (2021)
8. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: ICLR. OpenReview.net (2021)
9. Fumero, F., Alayón, S., Sánchez, J.L., Sigut, J.F., González-Hernández, M.: RIM-ONE: an open retinal image database for optic nerve evaluation. In: CBMS, pp. 1–6. IEEE Computer Society (2011)
10. Ilharco, G., et al.: Patching open-vocabulary models by interpolating weights. CoRR abs/2208.05592 (2022)
11. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D.P., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. In: UAI, pp. 876–885. AUAI Press (2018)
12. Jiang, M., Yang, H., Li, X., Liu, Q., Heng, PA., Dou, Q.: Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. MICCAI 2022. LNCS, vol. 13433, pp. 196–206. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_19
13. Kaddour, J., Liu, L., Silva, R., Kusner, M.J.: Questions for flat-minima optimization of modern neural networks. CoRR abs/2202.00661 (2022)
14. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: CVPR, pp. 10713–10722. Computer Vision Foundation/IEEE (2021)
15. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys. mlsys.org (2020)
16. Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S.: Multi-site FMRI analysis using privacy-preserving federated learning and domain adaptation: abide results. Med. Image Anal. **65**, 101765 (2020)
17. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: federated learning on non-IID features via local batch normalization. In: ICLR. OpenReview.net (2021)
18. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (2017)
19. Mirzadeh, S., Farajtabar, M., Görür, D., Pascanu, R., Ghasemzadeh, H.: Linear mode connectivity in multitask and continual learning. In: ICLR. OpenReview.net (2021)
20. Oh, J., Kim, S., Yun, S.: Fedbabu: Towards enhanced representation for federated image classification. CoRR abs/2106.06042 (2021)
21. Orlando, J.I., et al.: REFUGE challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Anal. **59**, 101570 (2020)

22. Pati, S., et al.: Federated learning enables big data for rare cancer boundary detection. Nat. Commun. **13**(1), 7346 (2022)
23. Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., Lu, Z.: Generalized federated learning via sharpness aware minimization. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 18250–18280. PMLR (2022)
24. Ramé, A., Ahuja, K., Zhang, J., Cord, M., Bottou, L., Lopez-Paz, D.: Recycling diverse models for out-of-distribution generalization. CoRR abs/2212.10445 (2022)
25. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**(1), 1–7 (2020)
26. Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al.: A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomed. Imaging Data Papers **2**(1), 1004 (2015)
27. Wortsman, M., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 23965–23998. PMLR (2022)
28. Wu, S., et al.: Motley: benchmarking heterogeneity and personalization in federated learning. CoRR abs/2206.09262 (2022)
29. Yao, Z., Gholami, A., Keutzer, K., Mahoney, M.W.: PyHessian: neural networks through the lens of the hessian. In: IEEE BigData, pp. 581–590. IEEE (2020)
30. Zhang, M., Sapra, K., Fidler, S., Yeung, S., Alvarez, J.M.: Personalized federated learning with first order model optimization. In: ICLR. OpenReview.net (2021)