Abstract—Federated learning is an emerging research paradigm for enabling collaboratively training deep learning models without sharing patient data. However, the data from different institutions are usually heterogeneous across institutions, which may reduce the performance of models trained using federated learning. In this study, we propose a novel heterogeneity-aware federated learning method, SplitAVG, to overcome the performance drops from data heterogeneity in federated learning. Unlike previous federated methods that require complex heuristic training or hyper parameter tuning, our SplitAVG leverages the simple network split and feature map concatenation strategies to encourage the federated model training an unbiased estimator of the target data distribution. We compare SplitAVG with seven state-of-the-art federated learning methods, using centrally hosted training data as the baseline on a suite of both synthetic and real-world federated datasets. We find that the performance of models trained using all the comparison federated learning methods degraded significantly with the increasing degrees of data heterogeneity. In contrast, SplitAVG method achieves comparable results to the baseline method under all heterogeneous settings, that it achieves 96.2% of the accuracy and 110.4% of the mean absolute error obtained by the baseline in a diabetic retinopathy binary classification dataset and a bone age prediction dataset, respectively, on highly heterogeneous data partitions. We conclude that SplitAVG method can effectively overcome the performance drops from variability in data distributions across institutions. Experimental results also show that SplitAVG can be adapted to different base convolutional neural networks (CNNs) and generalized to various types of medical imaging tasks. The code is publicly available at https://github.com/zm17943/SplitAVG.

Index Terms—Biomedical imaging, data heterogeneity, federated learning.

I. INTRODUCTION

Deep learning techniques and advances in computer hardware offer the promise of great advances in various medical applications, e.g., diagnosing ocular diseases [1], detecting glaucoma from optical coherence tomography images [2], identifying serious illnesses with natural language processing [3], and providing appropriate treatment recommendation [4]. However, training a robust deep learning model that generalizes across centers often requires a tremendous amount of training cases. The amount of patient data at individual medical institutions, or even in public data repositories such as The Cancer Imaging Archive, is often limited, especially for rarer diseases [5], [6]. Aggregating patient data from multiple centers is often complicated owing to patient privacy, legal regulatory barriers to data sharing, and inefficiency of moving large amounts of data. As such, federated learning (also termed as “collaborative learning” or “distributed learning”), where the training of a global deep learning model is performed locally at each institution without sharing raw data, has become a promising alternative for accessing large scale data to train robust deep learning models [7]–[14].

Existing federated learning methods can be grouped into aggregation-based federated learning methods [9] and transfer-based methods [15], [16]. Aggregation-based federated learning methods repeatedly average weights or gradient updates of the models trained at local institutions, such as Federated Averaging (FedAvg) and Federated stochastic gradient descent (FedSgd) [9]. Transfer-based methods train a model at each local institution for specific number of iterations and then transfer full or part of the model weights to next institution until model convergence, such as in Cyclic weight transfer (CWT) [15] and SplitNN [16] respectively.

One of the key challenges to federated learning algorithms is mitigating the effects of data heterogeneity among participating institutions on performance of the final learned model [17]. The distributed nature of federated learning means that there can be substantial heterogeneity in the distributions of training data across institutions. Aggregation-based methods like FedAvg can be robust to certain non-IID (independent and identically distributed) settings [9], but the accuracy of the synchronized averaged model reduces significantly on highly skewed data partitions [18]. Transfer-based methods may also lose model performance when heterogeneity exists in the training data across institutions, since the model trained in cyclic transferring
way always suffers from catastrophic forgetting on non-IID settings [19]. It has been demonstrated that training on non-IID data partitions is a pervasive problem for federated learning methods, and it always degrades the performance of deep learning models [21].

Several recent efforts have been devoted to overcoming the degrading effects of heterogeneous data across institutions in federated learning. The related studies include adding momentum to server model weight updates to prevent client updates trained on non-IID data partitions from diverging (FedAvgM) [22], applying the group normalization (Group Norm) [23] layers as the alternative of batch normalization to avoid the skew-induced accuracy loss of the batch normalization layer for non-IID data (FedSGD+GD) [21], [24], and sharing a subset of global data with local institutions to improve the training of FedAvg by mitigating weight divergence due to non-IID data (FedAVG+SD) [18]. Though these are promising approaches, the current approaches only work well on partitions with mild data distribution heterogeneity and still suffer performance drops on highly heterogeneous cases (see the comparison results in Fig. 5 for details).

In this study, we propose a novel federated learning technique, Split Averaging (SplitAVG), to overcome the deleterious effects of heterogeneous data distributions across institutions1 in federated learning. At the heart of SplitAVG is a network splitting operation and an intermediate feature map concatenation strategy. Specifically, our SplitAVG splits the network into an institutional sub-network (residing at the local institutions) and a server-based sub-network (residing on a central server) at a predefined layer of the network (see Fig. 1). As the training examples in each local institution are sampled from institution-specific data, which is a biased-estimator of the actual distribution of the whole population on non-IID data partitions, we further apply a concatenation operation on the central server to concatenate all the intermediate feature maps collecting from the institutional sub-networks. This concatenation operation allows SplitAVG to learn from the union of institution-specific data distribution rather than directly learning a biased-estimator of the actual distribution of the whole population, thus working well on both IID and non-IID data partitions. Our experimental results demonstrate the capability of the proposed method in handling unbalanced and non-IID data partitions.

The remainder of this paper is organized as follows: 1) present our SplitAVG algorithm, detail its forward propagation and back propagation training stage, 2) detail the binary classification dataset and regression datasets used to evaluate our method, 3) compare SplitAVG with seven state-of-the-art federated learning methods and the baseline centrally-hosted method, 4) present the experimental setup, and 5) provide an experimental evaluation of SplitAVG and its comparison methods on both IID data partitions and various non-IID data partitions.

II. MATERIALS AND METHODS

A. SplitAVG

In this section, we outline our proposed federated learning platform, SplitAVG (see Fig. 1), and provide a prospective to understand the advantage of the proposed method.

Define the deep learning network involved in SplitAVG as a function $F$, which consists of a list of $N$ sequential layers, i.e., $F = \{l_1, l_2, \ldots, l_c, \ldots, l_N\}$. In SplitAVG, we split $F$ into two sub-networks at a specific layer (also known as cut layer) $l_c$ and rewrite $F = \{F_I, F_S\}$, where $F_I = \{l_1, l_2, \ldots, l_c\}$ is the institutional sub-network that resides at the local institutions, and $F_S = \{l(c+1), l(c+2), \ldots, l_N\}$ is the server-based sub-network that resides on a central server. In each round of federated training, each local institution trains the institutional sub-network $F_I$ in parallel with its local data, sends the output feature maps to the central server for concatenating with those from other institutions, then the server completes the rest of the training with the aggregated feature maps on the server-based sub-network. Specifically, as depicted in Algorithm 1, SplitAVG follows a two-stage training phase: 1) data forward propagation procedure from institutional sub-network to the server-based sub-network with the transfer of the intermediate feature maps and their corresponding labels, 2) aggregated gradient back propagation procedure from server-based sub-network to the institutional sub-network. This two-stage training process is continued until model convergence on a separate validation set or the maximum number of training epochs is reached. Once the training is finished, the server will send the weights of server-based sub-network $F_S$ back to each local institution. Then each institution can perform validation and testing with the complete network $F = \{F_I, F_S\}$.

1) Forward Propagation: Let’s assume there are total $K$ local institutions involved in federated learning, indexed by $k$, and denote the training data of institution $k$ as $\{x_k, y_k\}$. In SplitAVG, we select a subset of $St \ll K$ local institutions following the client-selecting methods in FedAvg [9], and then start the

---

1The data heterogeneity in our study indicates heterogeneity from cross-institution data, also written as non-IID or data skew.
Algorithm 1: SplitAVG. The $K$ institutions are indexed by $k$; $E$ is the number of training epochs; $\{x_k, y_k\}$ is the data at institution $k$; $F = \{I(1), I(2), \ldots, I(c), \ldots, I(N)\}$ is the network layer, and $l$ is a layer in $F$, $lc$ is the cut layer. $\eta$ is the learning rate.

**Initialize:**

- $F = \{FI, FS\}$ \(\triangleright\) Institutions and the server split $F$ into institutional sub-network $FI$ and server-based sub-network $FS$
- $FI = \{I(1), I(2), \ldots, I(c)\}$
- $FS = \{(l(c + 1), l(c + 2), \ldots, lN)\}$ \(\triangleright\) $lc$ as the cut layer

Server initializes weight $W$ for $FI$ and sends $W$ to $K$ institutions

Server initializes weight $W_S$ for $FS$

for each epoch $i$ from 1 to $E$ do:

Server samples a subset of $St$ institutions

for each institution $k \in St$ do:

For each local batch $x_k, y_k$ in $x_k, y_k$ do:

**Forward Propagation:**

$x_{c,k}^l \leftarrow FI(x_k)$ \(\triangleright\) Institution $k$ forward propagates $x_k$ to cut layer $lc$

Server $\leftarrow x_{c,k}^{lc}, y_k$ \(\triangleright\) Institution $k$ sends intermediate feature maps at $lc$ and labels to server

$X_S^{lc} \leftarrow x_{c,1}^{lc} \oplus x_{c,2}^{lc} \ldots \oplus x_{c,N}^{lc}$ \(\triangleright\) Server concatenates feature maps from all institutions

$Y_S \leftarrow y_{1} \oplus y_{2} \ldots \oplus y_{St}$ \(\triangleright\) Server concatenates labels from all institutions

$Y_S \leftarrow FS(X_S^{lc})$ \(\triangleright\) Server forward propagates concatenated feature maps $X_S^{lc}$ till the final layer

**Back propagation:**

$g^N = \nabla L(Y_S, Y_S)$ \(\triangleright\) Server generates gradients at output layer

$g^l \leftarrow W_S^{T(l+1)}g^{l+1}$, for $l = l(N - 1), l(N - 2), \ldots, l(c + 1)$ \(\triangleright\) Server back propagates gradients to $l(c + 1)$

Institution $k \in St$ $\leftarrow g^{l(c+1)}$ \(\triangleright\) Server sends gradients at $l(c + 1)$ to each local institution

$g^l \leftarrow W_k^{T(l+1)}g^{l+1}$, for $l = lc, l(c - 1), \ldots, l1$ \(\triangleright\) Institution $k$ back propagates gradients

$W_k^l \leftarrow W_k^l - \eta g^l$, for $l = lc, l(c - 1), \ldots, l1$ \(\triangleright\) Institution $k$ updates $FI$

$W_S^l \leftarrow W_S^l - \eta g^l$, for $l = lN, l(N - 1), \ldots, l(c + 1)$ \(\triangleright\) Server updates $FS$

**Weight Transfer**

Institution $k \in 1, \ldots, K \leftarrow W_S$ \(\triangleright\) Server sends $FS$ to complete institution models

following forward propagation steps: 1) apply standard forward propagation on institutional sub-network $FI$ with sampled min-batch $\{x_k, y_k\}$ in each selected local institution, getting intermediate feature maps $FI(x_k)$; 2) send the intermediate feature maps and their corresponding labels $\{FI(x_k), y_k\}$ to the central server; 3) concatenate the received feature maps $X_S^{lc} = \{FI_1(x_1) \oplus FI_2(x_2) \ldots \oplus FI_{St}(x_{St})\}$ and their corresponding labels $Y_S = \{y_1 \oplus y_2 \ldots \oplus y_{St}\}$ at the central server; and finally 4) forward propagate the combined feature maps into the server-based sub-network $FS(X_S^{lc})$. This will complete a round of forward propagation without sharing the raw data. Unlike traditional federated learning methods, such as FedAvg [9], that directly average the model weights learned from institutional specific data distribution, where the synchronized averaged central model will lose accuracy or even completely diverge when high heterogeneity exists in the data partitions [18], our concatenation of feature maps on the central server guarantees that server-based sub-network is trained on the union of all the institutional data and not from biased institution-data, thus it works well on both IID and non-IID data partitions.

2) Back Propagation: After a round of forward propagation, SplitAVG back propagates the gradients from the last layer of the server-based sub-network to the first layer of the institutional sub-network $FI$. Given the loss function $L$, the detailed back propagation of SplitAVG is shown as the following: 1) calculate the gradients $g^N_S = \nabla L(Y_S, FS(X_S^{lc}))$ of the server-based sub-network $FS$; 2) back propagate gradients $g^N_S$ at the central server from the last layer of server-based sub-network $FS$ to its first layer, and denote the gradient at the first layer of server-based sub-network as $g^{l(c+1)}$; 3) transfer the gradients $g^{l(c+1)}$ back to each local institution and complete the rest of the back-propagation operation through each institutional sub-network $FI$; 4) update the model weights of both the server-based sub-network $FS$ and the institutional sub-network $FI$. Our back propagation procedure strictly follows the chain rule in differentiation, and it will achieve exactly the same results as the normal deep learning training procedure.

As opposed to traditional federated learning methods (such as FedAvg and FedSGD) [9] that require frequent transfers of model weights or model gradients of the entire network, in SplitAVG, only the intermediate feature maps $X_S^{lc}$ and gradients $g^{l(c+1)}$ at the cut layer are communicated between local institutions and the central server, which greatly reduces the computation and communication costs. In addition, the direct feature map concatenation step in the central server provides convergence guarantees for the model training and allows us to train a robust model on both IID and non-IID data partitions.

B. Theory Analysis for SplitAVG

Assuming there are $K$ local institutions. Let $T_k$ define the $k^{th}$ institution’s task domain where the raw data distribution is $D_{h_k}$, the local model learnt on this domain is $h_k$, and the empirical risk of the model on $T_k$ is $\mathcal{L}_{T_k}(h_k)$. Let $T$ denote the global domain where the data distribution $D$ including $m$ samples is assumed to be unbiased. $h$ denotes a global model achieved from federated learning which aims to minimize the risk of task from all $K$ local sites, written as $\mathcal{L}_T(h)$. According to the generalization bounds theory for federated learning in [25], $\mathcal{L}_T(h)$ has an upper bound that with probability larger than $1 - \delta$:

$$\mathcal{L}_T(h) \equiv \mathcal{L}_T\left(\frac{1}{K} \sum_k h_k\right)$$
where $d$ denotes the divergence measured between two domains, and $\tilde{D}_k$ and $\tilde{D}$ are intermediate feature representations reduced from raw images in $D_k$ and $D$ by a same feature extraction structure. The implication was derived that data heterogeneity of local sites to the global distribution leads to a high representation divergence $d(\tilde{D}_k, \tilde{D})$. The divergence increases the risk bound of the aggregated model $h$ on the global domain thus diminishing model quality. SplitAVG was motivated to reduce feature representation divergence, by proposing a concatenation operation on representations from selected local client:

$$\tilde{D}_k \leftarrow \{\tilde{D}_1 \oplus \tilde{D}_2 \ldots \oplus \tilde{D}_{St}\}$$

at the server sub-network to reduce the distance between the collection to the global feature distribution $\tilde{D}$. With this approach, SplitAVG lowers the upper bound of aggregated model $h$ on the global domain thus diminishing model quality. SplitAVG was motivated to reduce feature representation divergence, by proposing a concatenation operation on representations from selected local client: $\tilde{D}_k \leftarrow \{\tilde{D}_1 \oplus \tilde{D}_2 \ldots \oplus \tilde{D}_{St}\}$ at the server sub-network to reduce the distance between the collection to the global feature distribution $\tilde{D}$. With this approach, SplitAVG lowers the upper bound of aggregated model risk $L_T(h)$ without touching raw data space $D_k$. According to the bound, cut layer selection in SplitAVG does not affect the generalization performance of the final model if $d(\tilde{D}_k, \tilde{D})$ is determined, instead, it affects model’s learning ability of the server-based sub-network in extracting knowledge from $\tilde{D}$ and drawing hypothesis. This influence might differ across medical imaging tasks but follow the same empirical conclusion that the earlier cut layer is, the more parameters the server sub-network contains to facilitate learning.

C. SplitAVG-v2

The proposed SplitAVG algorithm includes the process of local institutions sending data labels to the central server, which brings in risks of privacy leaking, especially in tasks with high-dimensional labels. To this end, we introduce SplitAVG-v2, an improved version of SplitAVG by keeping the labels in local institutions to solve the privacy leakage concern raised from label sharing. As the architecture shown in Fig. 2, we further introduce a split point in the later part of the server network, where output predictions are split into chunks that each chunk of predictions is derived from an institution’s data. Prediction chunks are sent back to corresponding institutions and a scalar loss is computed with local data labels. The server collects institutional losses and generates the final loss and gradients.

SplitAVG-v2 does not require institutions to share raw data or raw labels, while retaining SplitAVG’s essence in generating unbiased gradients from collected loss. To illustrate, we take cross entropy loss as example. The traditional cross entropy in central server of SplitAVG is defined as:

$$L_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{C} t_{ic} \log(p_{ic})$$

where $N$ is the number of data points, $C$ is is the number of classes, $t_{ic}$ is the true label and $p_{ic}$ is the SoftMax probability of class $c$ at data point $i$.

In SplitAVG-v2, we defined an institutional cross-entropy:

$$L_{CE_k} = -\sum_{i=1}^{N_k} \sum_{c=1}^{C} t_{ic} \log(p_{ic})$$

where $N_k$ is the number of data points at the $k^{th}$ client. $L_{CE_k}$ is independently computed at each local client thus preserves the privacy of label $t_{ic}$. The server in SplitAVG-v2 then collects the institutional cross-entropy from all local institutions, resulting in the overall loss $L_{CEK}$ as:

$$L_{CEK} = \sum_{k=1}^{K} L_{CE_k}$$

$$= -\left[ \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{c=1}^{C} t_{ic} \log(p_{ic}) \right] = L_{CE}.$$
Fig. 3. Simulated data partitions on Retina and BoneAge datasets to simulate heterogeneity in data among 4 simulated institutions. Data partitions on Retina dataset with (a) $K-S=0$, (b) $K-S=0.40$, (c) $K-S=0.56$ and (d) $K-S=0.67$. Data partitions on BoneAge dataset with (e) $K-S=0.29$, (f) $K-S=0.59$, (g) $K-S=0.73$, (h) $K-S=0.97$.

Even though a split point is introduced in SplitAVG-v2, the overall loss of SplitAVG-v2 is the same to the SplitAVG. Thus SplitAVG-v2 and SplitAVG will have the same experimental results if a same experimental setting is used for both models. We further tested SplitAVG-v2 on split 4 of Retina dataset, and obtained the identical results, mean accuracy of 76.5%, with the SplitAVG result.

D. Dataset and Data Partitions

We evaluate our method on a set of both synthetic and real-world federated datasets, including the simulated federated datasets by artificially introducing data heterogeneity on a Diabetic Retinopathy (Retina) binary classification dataset [26] and a Bone Age (BoneAge) prediction dataset [27], and the real-world federated Brain Tumor Segmentation (BraTS 2017) segmentation dataset [28]–[30]:

The Retina dataset consists of 44,351 pairs of left and right eye color digital retinal fundus images obtained from the Kaggle Diabetic Retinopathy competition [26]. Each image is labeled on a scale of 0–4 based on the disease severity of diabetic retinopathy (DR), where 0 indicates no DR, and 1–4 represent mild, moderate, severe, and proliferative DR, respectively. We binarize the image labels to Healthy (scale 0) and Diseased (scale 2, 3 or 4) to simplify model training, and the mild DR (scale 1) images were excluded [15]. Furthermore, we only utilize left eye images to avoid the possible confusion from inconsistent correlation between disease presence in left/right eyes. The dataset is randomly sampled to create a training set of 6000 images, a validation set of 3000 images, and a testing set of 3000 images. The images are pre-processed following Ben Graham’s methods [31]: rescaled to a radius of 300, subtracting the local average color, image clipping for boundary removal, and resized to 256 x 256 resolution. Random cropping (to 224 x 224), random rotations (0, 90, 180, or 270 degrees) and horizontal flips were applied for data augmentation.

The BoneAge dataset consists of 14,236 pediatric hand radiographs obtained from the Kaggle Radiological Society of North America (RSNA) Bone Age competition [27]. Each image is labeled with the skeletal age provided by expert reviewers. The dataset is randomly sampled to create a training set of 4572 images, a validation set of 1000 images, and a testing set of 1000 images.

We simulate four “institutions” and create four kinds of data partitions for both the Retina and BoneAge datasets: one homogeneous data partition, and three heterogeneous data partitions with label distribution skew. The degree of label distribution skew is controlled by the fraction of non-IID data and is scaled by the mean Kolmogorov-Smirnov (K-S) statistic between every two institutions. Specifically, the K-S value being 0 indicates homogeneity and 1 indicates entirely different distributions. Fig. 3 depicts the detailed data partitions.

The BraTS dataset consists of magnetic resonance imaging (MRI) brain scans of gliomas collected from multiple institutions [28]–[30]. Each scan is manually labeled with segmentation annotations of tumor regions [28]–[30]. In our experiments, we focus on the segmentation for whole tumor region, and we only use high-grade glioblastoma (HGG) scans in T2 Fluid Attenuated Inversion Recovery (FLAIR) modality. We randomly select scans from 45 subjects as the testing set and the rest scans (120 subjects) as the training set. As a real-world federated dataset, BraTS includes common types of data heterogeneity, i.e., imaging acquisition skew (the scans are collected from ten institutions with different imaging equipment and protocols), label distribution skew and sample size distribution skew (one institution contributes 69 subjects while some institutions only contribute 4 or 5 subjects).

E. Comparison Methods

We compare our SplitAVG method with seven state-of-the-art federated learning methods including four traditional
methods: FedAvg [9], FedSGD [9], CWT [15], and SplitNN [16], and three optimized methods proposed for non-IID data: Federated stochastic gradient descent with group normalization (FedSGD+GN) [21], Federated averaging with server momentum (FedAvgM) [22], and Federated averaging with globally shared data (FedAvg+SD) [18]. We use the performance of a model trained with centrally hosted data as the baseline approach, termed as “centrally hosted”. This represents the ideal situation for training deep learning models since all data are centralized.

FedAvg is an aggregation-based method. For each epoch, local institutions conduct \( \frac{Q}{B} \) training iterations, then transfer model weights to a central server, which averages the weights and transfers the updated weights back to individual institutions [9]. \( Q_k \) is the quantity of training samples at institution \( k \), and \( B \) is the local mini-batch size.

FedSGD is a full-communication version of FedAvg. For each training iteration, local institutions transfer model gradients to central server, which generates weights updated from aggregated gradients and transfers updated weights back to individual institutions [9].

CWT is a transfer-based method. For each epoch, local institutions conduct \( \frac{Q}{B} \) training iterations, where \( Q \) is the quantity of training samples of the centrally hosted data, and cyclically transfers model weights to the next training institution until model convergence [15].

SplitNN is a transfer-based method. For each epoch, local institutions conduct \( \frac{Q}{B} \) training iterations with weights and gradients transferred between institutions and the server. Specifically, for each iteration: (1) a local institution forward propagates training data until the cut layer and transfers the outputs at the cut layer to a central server, (2) the server completes the rest of the training with the received output, (3) the server generates gradients, back propagates through the cut layer to the institution, and updates the model weights, and (4) the institution transfers model weights to the next training institution until model convergence [15]. Similar to SplitAVG, SplitNN also splits the whole network architecture into two parts, and involves frequent transfer of intermediate feature maps and gradients between the central server and local institutions. However, unlike SplitAVG that trains institutional sub-networks in parallel and uses an aggregation operation to concatenate the intermediate feature maps in the server, SplitNN directly uses a serial and cyclical transfer training mode in each local institution, which always suffers from catastrophic forgetting when data heterogeneity exists across institutions.

FedSGD+GD is an optimization method for FedSGD, which applies GroupNorm layers to avoid the skew-induced accuracy loss of batch normalization layer for non-IID data [21]. We set the number of groups in GroupNorm layers to 32.

FedAvgM is an optimization method for FedAvg, which applies a momentum optimizer on server to improve its robustness on non-IID data partitions [9], [22]. We set the momentum parameter to 0.9.

FedAvg+SD is an optimization method for FedAvg, which applies a data-sharing strategy to improve the training of FedAvg on non-IID data partitions [9], [18]. Specifically, 5% of the global data was distributed and globally shared between all local institutions.

F. Experimental Setup

We choose 34-layer residual network (ResNet34) pre-trained on ImageNet as the base network for all methods on Retina and BoneAge datasets [32], [33]. All methods are implemented in Pytorch and optimized using SGD [34]. The objective function for Retina classification task and BoneAge regression task is binary cross-entropy and L1-norm, respectively. We set the mini-batch size \( B \) to 32, the learning rate to 0.001 (scaled 0.1 every 40 epochs), the momentum coefficient to 0.9. Final models are evaluated by calculating the accuracy of testing data for the Retina classification task, and the mean absolute error (MAE) between true age values and predicting age values of testing data for the BoneAge regression task.

For BraTS segmentation task, we use U-Net as the base model and Dice Loss as the objective function [35], [36]. The final models are evaluated by Dice Similarity Coefficient (DSC) between the true and predicted boundaries [36].

The numbers of selected institutions involved in each round of federated learning for all the comparing methods are set to 4 (\( S_t = 4, K = 4 \)) for the synthetic datasets (Retina and BoneAge). To ensure that the communication and storage costs of SplitAVG does not increase with more institutions, we also set (\( S_t = 4, K = 10 \)) for BraTS dataset.

III. RESULTS

A. Cut Layer Selection for SplitAVG

According to our theory analysis, the selection of the affiliated cut layer for the institutional sub-network and server-based sub-network in SplitAVG does not affect the convergence speed of our model, but how many learn-able parameters the server sub-network contains affects model performance. We investigate the optimal cut layer for the base model ResNet34 and ResNet50 on BoneAge dataset with 1 homogeneous data partition and 1 heterogeneous data partition among 4 participating institutions. ResNet34 and ResNet50 consist of the following sequential layers: “conv1”, “bn1”, “relu”, “maxpool”, “layer1”, “layer2”, “layer3”, “layer4”, “avgpool”, and “fc”, which are tested as cut layers respectively [34]. The selection of cut layer affects the final model performance. However, the optimal cut layer selection is irrelevant to data partition or base network types, that earlier cut layers tend to result in better performance (Fig. 4). This observation is consistent with what we inferred in theory analysis that SplitAVG method can learn more abundant unbiased information when the feature maps from local institutions are concatenated at earlier layers. Also, the results show that deeper cut layers do not significantly compromise the model performance, especially as the base model complexity increases, comparing ResNet50 to ResNet34 results. The models fail only when setting last layer as cut layer, when the server network does not have sufficient model learnable parameters to interpret concatenated feature maps. Experimental results show that the ResNet models trained with “conv1” as the cut layer obtained
Fig. 4. For SplitAVG, when trained with different cut layers of ResNet34 (a) and ResNet50 (b) on a homogeneous split and a heterogeneous split of BoneAge dataset, the model performance on validation dataset is shown by the test mean absolute error (MAE).

C. Model Robustness for a Different Deep Learning Architecture

We replace the base model ResNet34 with MobileNet-v2 for all algorithms [37]. The architecture of MobileNet-v2 consists of a “Feature” structure including 16 “InvertedResidual” blocks, and a “Classifier” layer [38]. Following the cut layer empirical results in ResNet34, we set the first “InvertedResidual” block in MobileNet-V2 as the cut layer. The predicted MAE is used to evaluate SplitAVG, CWT, and FedAvg+SD on the BoneAge dataset when MobileNet-v2 is used as the base model. SplitAVG again demonstrates the best performance among all the compared methods. On the most skewed data partition (Split 4), SplitAVG achieves 104.7% of the MAE obtained by the baseline (Fig. 6).

D. SplitAVG on a Real-World Federated Dataset

We use the BraTS segmentation dataset to test SplitAVG’s robustness to the real-world data heterogeneity settings [28]–[30]. The BraTs dataset contains multi-modal magnetic resonance imaging (MRI) scans of 285 subjects with brain tumors. It is collected from 10 institutions with varying equipment and imaging protocols, thus resulting in heterogeneous data distributions among different clients, see Fig. 9 for four examples of images obtained from different institutions. Following [39], we test the performance of SplitAVG on the whole tumor volume segmentation task and adopt the FLAIR modality as the input, comparing with the centrally hosted baseline, CWT, and FedAvg+SD methods. We performed three trials with each method and take the mean of segmentation results across the 10 participating institutions. The model trained with data centrally hosted obtained the mean DSC result of 85.67%, and the model trained with CWT, FedAvg+SD, and SplitAVG obtained the mean DSC results of 79.72%, 83.16%, and 84.6%, as shown in Fig. 8.

E. Analyzing SplitAVG from Interpretation Perspective

We further visualize the latent space embedding of the features (the first “fc” layer of ResNet34) from the models trained with SplitAVG, three federated learning optimization methods, and the baseline centrally hosted training, to aid our understanding of different model’s robustness on heterogeneous data splits from interpretation perspective. We use Retina test dataset and draw features computed over samples of healthy label and diseased label with two different colors with UMAPs [39]. As shown in Fig. 7, the baseline UMAP presents the best clustering for same classes. Among UMAPs of comparing federated learning...
methods, SplitAVG shows the clearest separation for different classes, while the features of healthy and diseased shown from FedAvgM, FedAvg+SD, and FedSVG+GN are highly entangled. This experiment again demonstrates the superiority of SplitAVG on heterogenous data.

IV. DISCUSSION

Federated learning has emerged as an attractive paradigm for enabling collaboratively training deep learning models without sharing patient data. Although numerous federated learning approaches have been proposed, a critical aspect of existing federated learning methods is that they either assume the data are IID across institutions or only consider mild skewed non-IID data distribution. The performance of models trained using these federated learning methods degrades with increasing degrees of data heterogeneity. In this study, we develop a heterogeneity-aware optimization platform, SplitAVG, to address the challenge of data heterogeneity in federated learning methods.

We first evaluate our SplitAVG method on the simulated distributed data by artificially introducing various degrees of label distribution skew on the Retina binary classification dataset [26] and bone age prediction dataset [27] and compare it with seven state-of-the-art federated learning methods. We found that all the compared federated learning methods are vulnerable to label distribution skew. For Retina test dataset, the accuracy of models trained using FedAvg, FedSGD, CWT, and SplitNN, decreases from 74.3%, 77.8%, 78.3%, and 78.4% on data partitions with mild degree of label distribution skewness (K-S=0.40) to 51.5%, 50.5%, 50.9%, and 50.4% on data partitions with high degree of label distribution skewness (K-S=0.67), respectively. Even with complex heuristic parameters tuning (e.g., FedSGD+GN requires the extra pre-training on the model with GN layers [21], and FedAvgM includes the tuning for momentum parameters [22]) or with the risk of sharing partial raw data (FedAvg+SD) [18], the compared methods still suffer from severe performance drops on highly heterogeneous data partition. With the help of simple network splitting strategy and the concatenation operation of intermediate feature maps, SplitAVG, however, successfully mitigates model performance.
loss caused by the label distribution skew even in the extreme heterogeneous cases.

We then investigate whether SplitAVG method can handle other kinds of data heterogeneity besides label distribution skew and if it performs well in other kinds of deep learning tasks besides image classification and regression. Experimental results on a real-world BraTS segmentation dataset show that, even when tested with a mixture of various types of data heterogeneity (quantity skew, imaging acquisition skew, label distribution skew, etc), SplitAVG still achieves comparable performance to the baseline centrally hosted case. In contrast to previous methods that different optimizations are required for each type of data heterogeneity, for example cyclic weighted loss for tackling label heterogeneity and proportional local training for handling sample size heterogeneity [40], our SplitAVG is more scalable and that can more broadly address the challenge of data heterogeneity across centers in federated learning.

One limitation is that we only study federated learning method performances with statistical data heterogeneity. There are other sources of heterogeneity, such as device heterogeneity (e.g., computer hardware and communication speed variation) and behavior heterogeneity (e.g., institutions may join in or drop out training at any time), which is an important area for future work. One data privacy concern for the proposed SplitAVG is the risk of reconstructing raw images from shared feature maps of the cut layer, which can be prevented by integrating privacy protecting techniques such as secure multi-party computation (MPC) [41] and differential privacy [42], and future work can develop adjusted configurations combining these techniques.

V. CONCLUSION

In this paper, we have proposed SplitAVG, a heterogeneity-aware optimization platform that tackles fundamental and pervasive data heterogeneity problems inherent in federated learning. SplitAVG can be consumed as an off-the-shelf federated learning platform and provides immediate improvements, without any complex hyper-parameter tuning, training heuristic, or additional training/fine-tuning. SplitAVG is also model agnostic.
References

[1] H. Zhao, B. Yang, L. Cao, and H. Li, “Data-driven enhancement of blurry retinal images via generative adversarial networks,” in Medical Image Computing and Computer Assisted Intervention - MICCAI. Berlin, Germany: Springer, 2019, pp. 75–83.

[2] X. Wang et al., “Unifying structure analysis and surrogate-driven function regression for glaucoma OCT image screening,” in Medical Image Computing and Computer Assisted Intervention - MICCAI. Berlin, Germany: Springer, 2019, pp. 39–47.

[3] B. Udelsman, I. Chien, K. Ouchi, K. Brizzi, J. A. Tulsy, and C. Lindvall, “Needle in a haystack: Natural language processing to identify serious illness,” J. Palliat. Med., vol. 22, no. 2, pp. 179–182, Feb. 2019.

[4] D.-K Hwang et al., “Artificial intelligence-based decision-making for age-related macular degeneration,” Theranostics, vol. 9, no. 1, pp. 232–245, Jan. 2019.

[5] V. Tresp, J. M. Overhage, M. Bundschus, S. Rabizadeh, P. A. Fasching, and S. Yu, “Going digital: A survey on digitalization and Large-Scale data analytics in healthcare,” Proc. IEEE, vol. 104, no. 11, pp. 2210–2206, Nov. 2016.

[6] K. Clark et al., “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” J. Digit. Imag., vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[7] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proc. 22nd ACM SIGSAC, Conf. Computer Commun. Secur., 2015, pp. 1310–1321.

[8] A. Jochems et al., “Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept,” Radiother. Oncol., vol. 121, no. 3, pp. 459–467, Dec. 2016.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. 20th Int. Conf. Artif. Intell. Statist., 2017, pp. 1273–1282.

[10] G. Elmas et al., “Federated learning of generative image priors for MRI reconstruction,” 2022, arXiv:2202.04175.

[11] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K.-T. Cheng, “Variation-aware federated learning with multi-source decentralized medical image data,” IEEE J. Biomed. Health Inform., vol. 25, no. 7, pp. 2615–2628, Jul. 2021.

[12] Q. Liu, C. Chen, J. Qin, Q. Dou, and P. Heng, “FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1013–1023.

[13] L. Qu et al., “Rethinking architecture design for tackling data heterogeneity in federated learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 10061–10071.

[14] L. Qu, N. Balachandar, M. Zhang, and D. Rubin, “Handling data heterogeneity with generative replay in collaborative learning for medical imaging,” Med. Image Anal., vol. 78, 2022, Art. no. 102424.

[15] K. Chang et al., “Distributed deep learning networks among institutions for medical imaging,” J. Amer. Med. Inform. Assoc., vol. 25, no. 8, pp. 945–954, Aug. 2018.

[16] O. Gupta and R. Raskar, “Distributed learning of deep neural network over multiple agents,” J. New. Comput. Appl., vol. 116, pp. 1–8, Aug. 2018.

[17] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning From Imbalanced Data Sets. Berlin, Germany: Springer, 2018.

[18] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civan, and V. Chandra, “Federated learning with Non-IID data,” J. 2018, arXiv:1806.05952.

[19] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, “Overcoming catastrophic forgetting by incremental moment matching,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 4655–4665.

[20] L. Qu, N. Balachandar, M. Zhang, and D. Rubin, “Handling data heterogeneity with generative replay in collaborative learning for medical imaging,” Med. Image Anal., vol. 78, 2022, Art. no. 102424.

[21] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, “The Non-IID data quagmire of decentralized machine learning,” in Proc. 37th Int. Conf. Mach. Learn., 2020, pp. 4387–4398.

[22] T.-M. H. Hsu, H. Qi, and M. Brown, “Measuring the effects of Non-Identical data distribution for federated visual classification,” Sep. 2019, arXiv:1909.06335.

[23] Y. Wu and K. He, “Group normalization,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 742–755.

[24] L. G. Valiant, “A bridging model for parallel computation,” Commun. ACM, vol. 33, no. 8, pp. 103–111, Aug. 1990.

[25] Z. Zhu, J. Hong, and J. Zhou, “Data-free knowledge distillation for heterogeneous federated learning,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 12878–12898.

[26] Kaggle, 2017. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[27] S. S. Hallabi et al., “The RSNA pediatric bone age machine learning challenge,” Radiol., vol. 290, no. 2, pp. 498–503, Feb. 2019.

[28] B. H. Menze et al., “The multimodal brain tumor image segmentation benchmark (BRATS).” IEEE Trans. Med. Imag., vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[29] S. Bakas et al., “Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features,” Sci. Data, vol. 4, Sep. 2017, Art. no. 170117.

[30] E. Carver et al., “Automatic brain tumor segmentation and overall survival prediction using machine learning algorithms,” in Proc. Int. MICCAI Brainstesion Workshop, 2019, pp. 406–416.

[31] B. Graham, “Kaggle diabetic retinopathy detection competition report,” Univ. Warwick, pp. 24–26, 2015.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[33] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[34] S. Ruder, “An overview of gradient descent optimization algorithms,” 2016, arXiv:1609.04747.

[35] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” Med. Image Comput. Assist. Intervention, pp. 234–241, 2015.

[36] L. R. Dice, “Measures of the amount of ecologic association between species,” Ecology, vol. 26, no. 3, pp. 297–302, 1945.

[37] M. Sandler et al., “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[39] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, “Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation,” Brainlesion, vol. 11383, pp. 92–104, Jan. 2019.

[40] N. Balachandar, K. Chang, J. Kalpathy-Cramer, and D. L. Rubin, “Accounting for data variability in multi-institutional distributed deep learning for medical imaging,” J. Amer. Med. Inform. Assoc., vol. 27, no. 5, pp. 700–708, May 2020.

[41] O. Goldreich, “Secure multi-party computation,” Manuscript Preliminary version, vol. 78, 1998, Art. no. 110.

[42] M. Abadi et al., “Deep learning with differential privacy,” in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2016, pp. 308–318.