Decentralized Sparse Federated Learning for Efficient Training on Distributed NeuroImaging Data

Bishal Thapaliya[1,3], Riyasat Ohib[2,3], Jingyu Liu[1,3], Vince Calhoun[1,2,3], Sergey Plis[1,3]
– [1]Georgia State University, [2] Georgia Institute of Technology, [3]TReNDs Center

TReNDS Center for Translational Research in Neuroimaging & Data Science · Georgia Tech · Georgia State University

Summary

Neuroimaging advancements have increased data sharing among researchers. Yet, institutions often retain data control due to research culture, privacy, and accountability. There is therefore a need for tools that analyze combined datasets without transmitting the actual data. We introduce a decentralized sparse federated learning (FL) approach that locally trains sparse models for efficient communication in such settings. By leveraging sparsity and transmitting only some parameters among client sites throughout the training, we reduce communication costs, especially with larger models and varied site-specific resources. We validate our method using the ABCD data.

Background

Gradient-based measures [1] identify important connections in the network by utilizing a sensitivity measure defined as the change in the loss of the connections in question were removed. More formally, the effect of the weight $\boldsymbol{\theta}_j$ on the loss is:

$$s(\theta_j) = |\theta_j \odot g_j| \tag{1}$$

In the FL setting for N different client sites using the saliency criterion in 1, we get the saliency score $s(\theta_0; D_k)$ for the k^{th} site. To calculate the score S_k we pass a few minibatches of data and average the saliency scores over the few minibatches. To create the global mask m_g we average all the saliency scores from N different sites and apply the *top-k* operator to find the most important connections based on the initialization and the data on all the client sites. Thus, to generate the global mask $\mathbf{m}_g \in \{0,1\}^{|\boldsymbol{\theta}|}$ we select for the *top-k* ranked connections as:

$$m_g = \mathbb{1}[s(\theta_j) \geq s(\theta_k)]$$

where, θ_k is the k^{th} largest parameter in the model and $\mathbb{1}[\cdot]$ is the indicator function.

Our study focuses on the task of classifying a participant's sex based on MRI scans, by employing a 3D variant of the well-known AlexNet model. The 3D variant was specific channel configuration for the convolutional layers set as: 64C-128C-192C-192C-128C, where 'C' denotes channels.Our training consists of 5 epochs with 200 communication rounds.

Contributions

Sparse Salient Federated Learning (SSFL) is a novel paradigm for federated learning with sparse models that aims to address issues of current sparse FL methods. The primary benefit of our approach is that we find sparse models to be trained using the information from the data at all the client sites and find a sparse network or a subnetwork to be trained *before* the training begins.

- We train only a subset of the parameters of client models in a decentralized manner, resulting in a highly communication-efficient federated training technique, resulting in sparse models at client sites.
- We propose a generalized version of gradient-based connection importance criterion for Federated Learning in the non-IID setting.
- *SSFL* identifies a subnetwork prior to training, leveraging parameter saliency scores keeping in mind the distribution of local client data in non-IID scenarios

Algorithm

Algorithm 1 SSFL

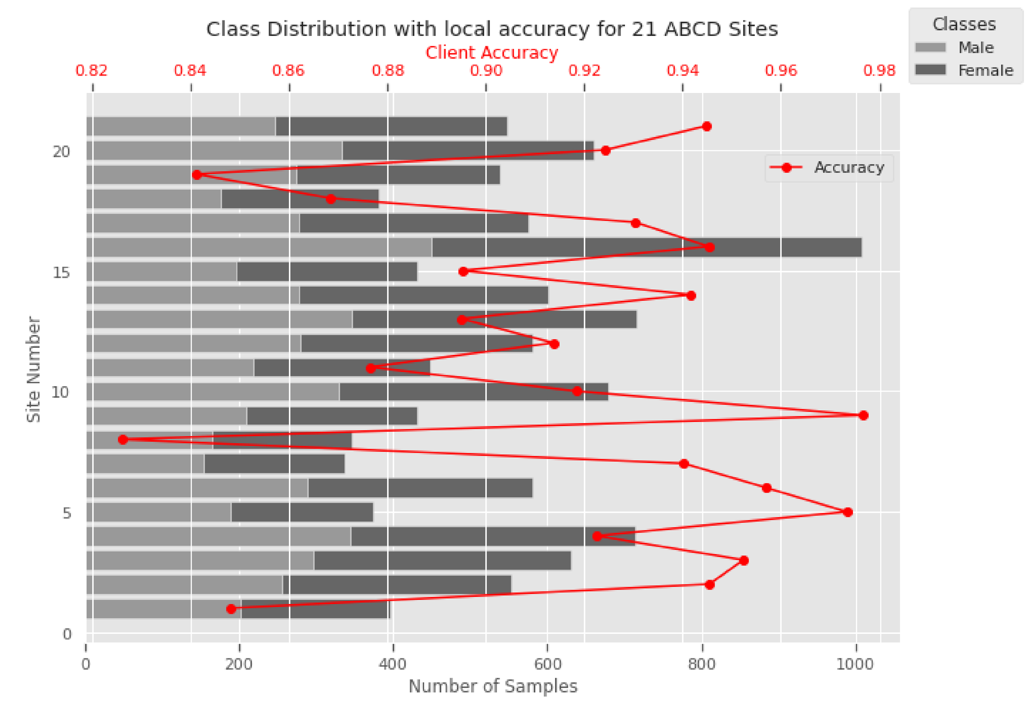
Input: Total number of clients K; Total communication rounds R; Total local training steps T in each communication round; Number of clients K' participating in each round.
Output: Sparse local models $\hat{w}_m^{\mathcal{C}}$

```
Initialize model with parameters w_0 and transmit to all clients.
Calculate proportion of data at each client p_k.
s = Σ_{k=1}^{K} p_k s_k(w_0) # Get the aggregated global saliency scores
m = 1[s(w_j) > s(w_k)] # calculate common mask from aggregated saliencies
Transmit m to all the sites.
w_{k,m} ← w_{k,0} ⊙ m  #apply the mask at all sites k

for r = 0 to R − 1 do
    c_1, c_2, ..., c_K̂ ∼ Unif(C) # Sample K̂ clients uniformly from the set of all clients C
    for site k in parallel for all K̂ clients do
        w_m ← csr(w_{k,m}); #Gather all masked weights w_{k,m} where k ∈ {1, 2, 3, ... K̂}
        w_m^Ĉ ← ( (1/K̂) Σ_{k=1}^{K̂} w_{k,m} )  #combine the weights of models in the selected sites
        ŵ_m^0 ← w_m^Ĉ

        for t = 0 to T − 1 do
            g_m^t ← ∇_w L(ŵ_m^t; x^t, y^t) ⊙ m # calculate and mask gradients
            ŵ_m^{t+1} ← ŵ^t_m − η g_m^t # take optimization step with masked gradients on masked weights
        end for
    end for
    transmit the non-zero elements of global models ŵ_m^C back to the clients.
end for
```

Results: SSFL vs DistPFL vs FedAvg-FT



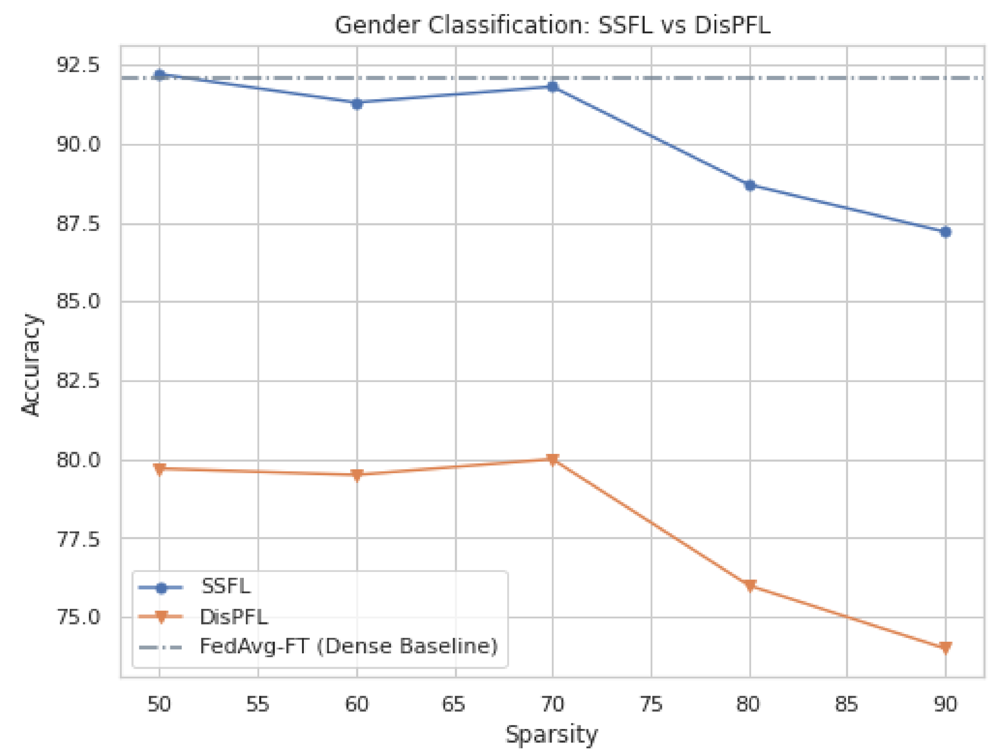


Figure (top): : Gender differences in each of the 21 ABCD sites along with the performance of the model. **Figure (bottom):** : Performance comparison of *SSFL* with *DistPFL* and *FedAvg-FT* on 3D AlexNet Model.

Conclusion

We propose *SSFL*, a novel federated learning paradigm that collaboratively trains a highly sparse model in a non-IID setting. The major benefits of SSFL over the existing methods 1) In contrast to most existing methods *SSFL*, only need to transmit highly sparse parameters between server and clients which significantly reduces the communication time as well as the bandwidth. 2) We compute a parameter saliency that captures the local data characteristics at client sites and create a global model mask based on that saliency, resulting in a client data-aware sparse model.

Reference

References

[1] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. "Snip: Single-shot network pruning based on connection sensitivity". In: *arXiv preprint arXiv:1810.02340* (2018).