Communication-efficient federated learning for multi-institutional medical image classification

Zhou, Shuang, Landman, Bennett, Huo, Yuankai, Gokhale, Aniruddha

Shuang Zhou, Bennett A. Landman, Yuankai Huo, Aniruddha Gokhale, "Communication-efficient federated learning for multi-institutional medical image classification," Proc. SPIE 12037, Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, 1203703 (4 April 2022); doi: 10.1117/12.2611654

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Communication-Efficient Federated Learning for Multi-Institutional Medical Image Classification

Shuang Zhou[a,c], Bennett A. Landman[b,a], Yuankai Huo[a,b], and Aniruddha Gokhale[a,b,c]

[a]Department of Computer Science, Vanderbilt University, Nashville, TN, 37235 USA
[b]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, 37235 USA
[c]Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, 37212 USA

ABSTRACT

Federated learning (FL) has emerged with increasing popularity in the medical image analysis field. In collaborative model training, it provides a privacy-preserving scheme by keeping data localized. In FL frameworks, instead of collecting data from clients, the server learns a global model by aggregating local training models from clients and broadcasts the updated model. However, in the situation where data is not identically and independently distributed (non-i.i.d), the model aggregation requires frequent message passing, which may face the communication bottleneck. In this paper, we propose a communication-efficient FL framework based on the adaptive server-client model transmission. The local model in the client will only be uploaded to the server under the conditions of (1) a probability threshold and (2) an informative model updating threshold. Our framework also tackles the data heterogeneity in federated networks by involving a proximal term. We evaluate our approach on a simulated multi-site medical image dataset for diabetic retinopathy (DR) rating. We demonstrate that our framework not only maintains the accuracy on non-i.i.d dataset but also provides a significant reduction in communication cost compared to other FL algorithms.

Keywords: Federated Learning, Medical Image Classification, Communication Efficiency

1. INTRODUCTION

The raw data in medical institutes, e.g., medical images and records, is quite sensitive to be exposed to other parties. Traditional machine learning algorithms, which require aggregating the distributed datasets at a central server, may incur practical challenges in collaborative model training. Federated learning (FL)[1] has recently emerged as a privacy-preserving solution by eliminating the potential to exchange sensitive data which is beneficial to the model training using medical data. The data privacy can be preserved by only transmitting model parameters between the server and clients, which avoids data violation across the sites.

However, FL would suffer from performance decreasing at clinical deployment. First, the datasets which are distributed across multi-sites inevitably fall in data heterogeneity. The variance of the collaborative datasets lacks assurance for good generalizability when facing complex medical data. Second, each client's local training can only learn its generalizable parameters from its isolated data distribution. The limitation of the data accessibility, which constrains the total usage of data distributions across institutes, may mislead the model training.

Also, in the FL scheme, a large number of institutes locate at different places, and attempt to communicate to the central server with its local updates. The communication constraints (i.e., bandwidth bottleneck) cannot be ignored. To achieve a higher accuracy, the size of the model becomes larger than before.[2] Therefore, the network requires considerable communication resources and faces a significantly longer transmission latency in uploading, which may lead to asynchronization in training stage.

Prior works concentrate on the model compression such as quantization[3] to reduce the communication overload, while we focus on reducing the transmission rate instead. We propose a novel approach in FL to reduce the

Further author information: Shuang Zhou: E-mail: shuang.zhou@vanderbilt.edu

Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications,
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 12037,
1203703 · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2611654

communication cost between clients and server by permitting the informative update under a certain probability. We also introduce a proximal term following the techniques in FedProx[4] to tackle the data heterogeneity.

In conclusion, our contributions are as follows:

- Our proposed approach presents a communication-efficient strategy by allowing the clients to determine whether to upload its model or not. Using this strategy, our model also provides a better performance when solving the problem of non-i.i.d.
- The experiments on the DR dataset[5] show that the communication cost is decreased by an average of 25% without significantly sacrificing the accuracy. We also demonstrate that our approach improves the testing accuracy by 2% on average and decreases the training loss by 28% in the highly heterogeneous dataset.

2. METHOD

We introduce a strategy with the goal of decreasing communication costs and improving accuracy. The proposed approach requires two major modifications of the original FedAvg. In particular, we add a proximal term to optimize the non-i.i.d problem and present a scheme where each client only transmits its update back to the server only if it meets an adaptive threshold.

2.1 Statistical Heterogeneity Optimization

In the Federated averaging (FedAvg)[6] , which is a basic yet effective algorithm for federated learning, a central server first distributes the initial global model to the clients (institutes) and selects a small fraction C of K clients to begin a new epoch of local training – where K is the number of total clients in the network. Then in each global iteration, the following server-client communication strategy with two stages will be repeated until convergence:

- The clients will be involved in collaboration by independently performing E epochs of training the global model on their local datasets with stochastic gradient descent (SGD) optimizer. The clients will reply to the server with the current round updated model.
- After collecting from all participated clients, the updated models will be aggregated in the server by averaging the updates with weights proportional to the size of the local dataset. Finally, the server starts a new round of broadcasting the latest global model to another set of selected clients.

However, statistical heterogeneity appears when data is non-identically distributed across the network. FedAvg does not fully address this underlying challenge. Our approach modified from FedAvg aims to keep the updated parameters across clients more similar to solve the non-i.i.d problem.

The goal of FL is to minimize the following distributed optimization model(1):

$$\min_{w} \left\{ \mathcal{L}(w) = \sum_{k=1}^{K} p^k \mathcal{L}^k(w) \right\}, \tag{1}$$

In the equation above, K is the number of clients in the network, and p^k is the proportional weight of the kth client such that $p^k \geq 0$ and $\sum_{k=1}^{K} p^k = 1$. $\mathcal{L}(w)$ is the user-specified loss function. In each round t, the updated model weight w_{t+1}^k at client k will be calculated as $w_{t+1}^k = w_t - \eta \bigtriangledown \mathcal{L}^k(w_t)$. The loss function $\mathcal{L}_k(w)$ at client k is defined by $l(w, n^k)$ given the n^k data and weight w .

Due to data heterogeneity across medical institute, the global model will derive its convergence bound after more epochs. Our approach, which is inspired by the FedProx,[4] tackles heterogeneity in federated networks by introducing a proximal term in the loss function calculation. In our approach, instead of minimizing the loss in (1), during the local update, client k will find w to minimize (2):

$$\min_{w} h^k(w, w_t) = \mathcal{L}^k(w) + \frac{\mu}{2} \left\| w - w_t \right\|^2 \tag{2}$$

Due to heterogeneity in data distributions, the divergence of parameters between clients will be increased. By adding the proximal term $\frac{\mu}{2} \left\| w - w_t \right\|^2$, $\mathcal{L}^k(w)$, the potential impact of various local updates, will be limited.

2.2 Communication-efficient strategy

Algorithm 1: Communication-Efficient Scheme

Input: K clients, the fraction c for selection, number of local rounds N, number of global epochs E, learning rate η, pre-defined threshold τ, probability p, proportional weights v for each client

Output: Global Model

Initialize global model θ;
for t = *1:* E **do**
  Server selects a subset C_t of K clients at random. The size of C_t is $c * K$;
  for k *in* C_t *in parallel clients* **do**
    Initialize $\widehat{w}_{t+1}^k \leftarrow \theta_t$, $o_{t+1}^k \leftarrow \tau_t^k$;
    for i = *1, ..., N* **do**
      Find a w_{t+1}^k to $\min_w h^k(w_{t+1}^k, w_t) = \mathcal{L}^k(w_{t+1}^k) + \frac{\mu}{2}\left\|w_{t+1}^k - w_t\right\|^2$;
      Update $w_{t+1}^k \leftarrow \widehat{w}_{t+1}^k$;
      $o_{t+1}^k = \left\|w_t^{k+1} - w_t^k\right\|_2$;
      if *(*$o_{t+1}^k < \tau_k$&p > 0.5*)* **then**
        Return (o_{t+1}^k , NONE) to server;
      else
        Return (o_{t+1}^k, w_t^{k+1}) to server
      end
    end
  end
  Server Updating: ;
  if *Receive all feedback* o_{t+1}^k *and* w_t^{k+1} *from* C_t **then**
    for k *in* C_t **do**
      Load state dictionary w_{t+1}^k to θ_{t+1}^k;
    end
    $\tau_{t+1} \leftarrow \sum_{k=1}^K v_k * o_{t+1}^k$;
    $\theta_{t+1} \leftarrow \sum_{k=1}^K v_k * \theta_t^{k+1}$;
  end
end

In a FL framework, larger networks will result in a concern of communication resources such as network uploading-downloading pipeline and bandwidth. To achieve the goal of efficient communication, existing approaches typically introduce compression of the updated model, such as subsampling model parameters[1] and quantization[37] on the client-side training. However, compression methods may bring along with additional computation for encoding and decoding[7] , which increase the local training duration. Also, FedPAQ,[3] a quantized message passing strategy, assumes the datasets in all clients are independent and identically distributed (i.i.d.), which cannot provide the same accuracy when processing the non-i.i.d medical data.

Our proposed method reduces the transmission load by employing two modules: partial node (client) participation and conditional clients update. We formalize our framework for communication-efficient FL as Algorithm 1.

During each global epoch t, the server will distribute the aggregated global model θ_t to the selected fraction C of total K clients. After receiving the model θ_t, client k will perform a local update on the global estimated weight w_t from θ_t by finding γ-inexact minimizer w_{t+1}^k of $h^k(w_{t+1}^k, w_t)$ in (2). By computing the $l2$ norm o_{t+1}^k of the model difference between w_t^k and w_{t+1}^k, w_{t+1}^k can be considered as informative update if o_{t+1}^k is larger than the an adaptive threshold τ_t. The threshold τ_t will be determined at the beginning of each epoch t by calculating the mean of the norms of uploaded o_t from every client.

Since the τ_t varies in different epoch, we will meet the challenge that most of the o_{t+1} in the chosen clients

below the τ_t during some particular epochs. If client k transmits the weight back to the server only depending on the result of comparison between o_{t+1}^k and τ_t, the global model will reach the convergence while resulting in significant accuracy loss. So we add a probability p as a conditional parameter to decide whether to perform the updating or not. In such situation, some of the updates will not be ignored even if they are considered as relatively less informative. Specifically, if o_{t+1} is smaller than τ_t, the clients will generate a random value p_i between(0,1). If p_i is larger than p, instead of transmitting the model back, the client i will transmit a "NONE" message which is only one float 32 bits along with the l2 norm o_{t+1}^k. It will significantly decrease the message size compared to the full model. Finally, the server computes the new global model θ_{t+1} by aggregating the received weights from $C * N$ clients with records of the weights for the remained clients (3):

$$\theta_{t+1} = \sum_{k=1}^{K} v^k * \widehat{\theta}_{t+1}, \quad \widehat{\theta}_{t+1} = \begin{cases} \theta_{t+1}^k, & \text{if received an update local from client t} \\ \theta_t^k, & \text{otherwise} \end{cases}, \tag{3}$$

where $v^k = \frac{n^k}{\sum_{i=k}^{K} n^i}$ is the weight of data set n^k.

Finally, at the server side, after receiving all the updates from the selected clients, the aggregation which processes threshold τ_{t+1} and global model θ_{t+1} for the next epoch $t+1$ is shown in Line 20-21 in Algorithm 1.

3. EXPERIMENTS

3.1 Dataset

Diabetic retinopathy is a diabetes complication which can cause vision loss. The diagnosis is made by examining scans of the blood vessels of the light-sensitive tissue at the back of the retina which turns the diagnosis into a medical image classification problem. A DR dataset[5] provides a real-world medical image analysis datasets with rated image for the severity. DR consists of 2931 variable-sized images for training, and 731 images for testing. The problem is to classify the images of patients' retina into five scalability categories from 0 to 4: 0 - No DR, 1 - Mild DR, 2 -Moderate DR, 3 - Severe DR, and 4 - Proliferative DR.

Like previous studies[8],[9] we use Dirichlet distribution to generate non-i.i.d datasets among batches. For each class j, we sample $p_j \sim Dir_K(\beta)$ where $Dir_K(\beta)$ is Dirichlet distribution and β is the concentration parameter. We allocate $p_{j,k}$ proportion of the instances of class j of the complete dataset to batch j and due to the small value of β, some batches may lack of an entire subset of classes. The data distributions among batches in different size are shown in Figure 1.

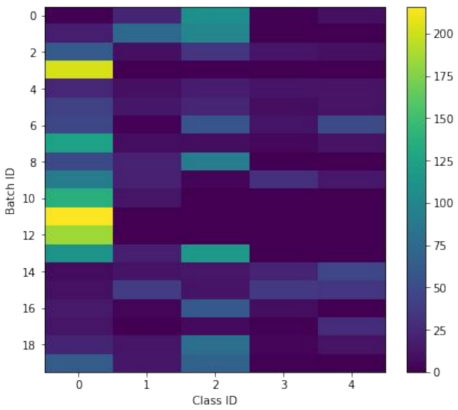

Figure 1. The data distribution of each party using non-IID data partition. The color bar denotes the number of data samples. Each rectangle represents the number of data samples of a specific class in a batch.

3.2 Implementation

One of the motivations of using SqueezeNet[10] is the less communication load compared to other famous models. In our experience, SqueezeNet not only has a small size of 4.73MB but also achieves a classification accuracy of about 80% on the DR dataset in the centralized training.[11]

We implement FedAvg, FedProx, and our strategy in PyTorch.[12] We use the SGD optimizer with a learning rate of 0.002 for all approaches to draw a fair comparison. For the heterogeneous dataset partition, we choose $\beta = 0.5$ to split the dataset for each of the classes into 20 unequal parts to form 20 batches (clients). The number of local rounds is set to 10, and the number of the global communication epochs is set to 100 for all federated learning approaches. The probability threshold p for conditional updating is set to 0.5, and the adaptive l2 norm threshold is initialized to be 5.

To demonstrate the performance of our proposed algorithm, we study our FL systems described above in regards to the following aspects: (i) the performance of communication efficiency and testing accuracy in the client selection strategy; (ii) the effect of proximal term in heterogeneous settings.

We first test our proposed client selection algorithm and compare the communication volume saving with the non-restricted communication scheme. All of the client selection at the server-side has an equal fraction of 0.5. We observe similar relative accuracy behavior of our strategy corresponding to the FedProx and better accuracy behavior compared to the FedAvg. Then we turn our attention to the policy for transmitting model updates to the server under a certain probability even though not exceeding the pre-defined threshold. We compute the l2 norm value during the total 100 global epochs. The update scheme herein prevents the risk of misjudgment of informativeness.

4. RESULTS

We first test our aggregation scheme with different values of proximal term. When $\mu = 0$, the strategy can be considered as the FedAvg with communication saving methods. In Figure 2, we observe that by adding the proximal term, the scheme improves the highest testing accuracy by 2% and decreases average training by 28% under the selected updating situation.

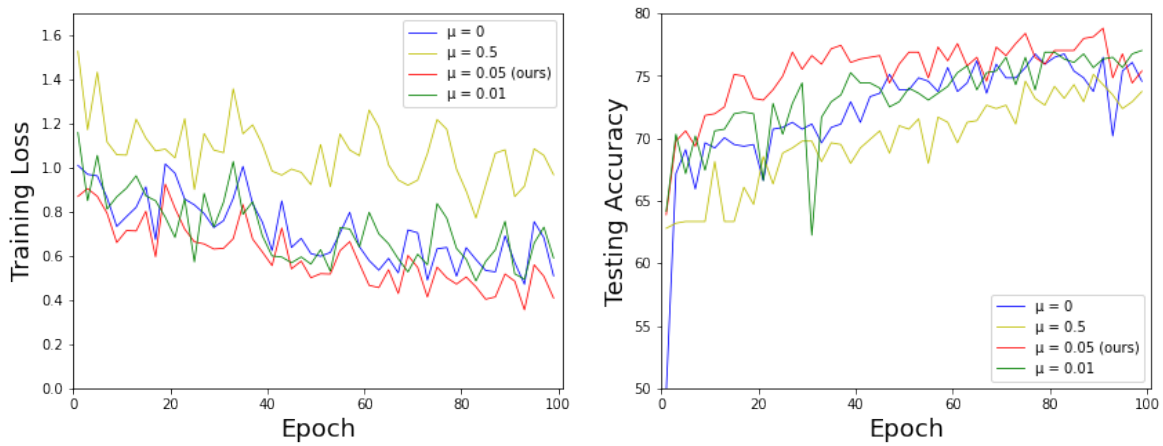

Figure 2. Training loss (left) and testing accuracy (right) between scheme with proximal term $\mu = 0.05, 0.1$ and FedAvg $\mu = 0$. Algorithm with proximal terms results in a better model accuracy and lower training loss compared to FedAvg.

As described in Section 3, to reduce communication consumption, we propose a strategy in which only the informative updates will be processes under a certain probability by compared between threshold τ_t and l2 norm which is the difference between the previous and the current weights in each round. In Figure 3, we notice that our approach provides nearly 25% lower communication cost compared to the approach which maintains the full communication such as FedProx while keeping the approximately similar accuracy performance.

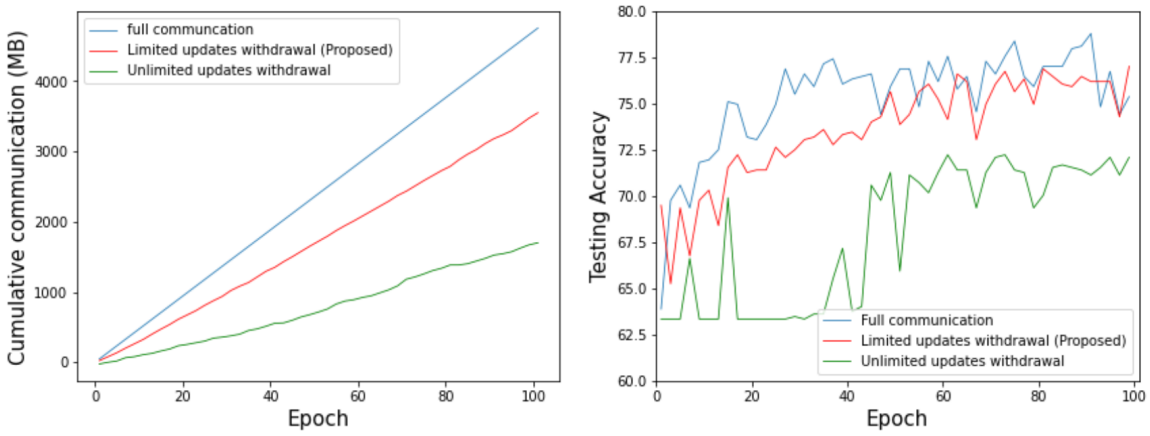

Figure 3. Cumulative communication (left) measured in MBs and testing accuracy (right) of full communication, limited updates withdraw, and unlimited updates withdraw during the training. The proposed method saves 25% transmitting cost without significant decrease of the accuracy compared to the full communication.

It is worth considering if the probability item p is eliminated which might save more rounds of communication overloads. Specifically, we observe that although the strategy without probability p achieves the nearly twice communication savings, it is not capable of matching the same level of accuracy.

5. CONCLUSIONS AND DISCUSSION

In this paper, we propose a novel communication-efficient FL method with favorable performance. Our approach consists of two modules: (1) tackling heterogeneity by adding a proximal term to the loss function; (2) selective model updating under the conditions of a probability threshold and an informative model determination threshold. We simulate various communication scenarios on a DR dataset with non-i.i.d settings and show that our approach significantly reduces the communication overhead while maintaining a satisfactory accuracy.

For future work, a promising direction is to study the challenge where the clients will join or leave the FL network without any notification. The records of all clients need to be traced for the whole training rounds. We will add a frequent monitoring strategy without significant communication costs.

6. ACKNOWLEDGMENTS

This work has not been submitted for publication or presentation elsewhere. This research was supported by NSF CAREER 1452485, NSF Convergence Accelerator Track D 2040462, and R01 EB017230.

REFERENCES

[1] Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D., "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492* (2016).

[2] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V., "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582* (2018).

[3] Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R., "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in [*International Conference on Artificial Intelligence and Statistics*], 2021–2031, PMLR (2020).

[4] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V., "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127* (2018).

[5] Choi, J. Y., Yoo, T. K., Seo, J. G., Kwak, J., Um, T. T., and Rim, T. H., "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database," *PloS one* **12**(11), e0187336 (2017).

[6] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A., "Communication-efficient learning of deep networks from decentralized data," in [*Artificial intelligence and statistics*], 1273–1282, PMLR (2017).

[7] Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M., "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems* **30**, 1709–1720 (2017).

[8] Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y., "Bayesian nonparametric federated learning of neural networks," in [*International Conference on Machine Learning*], 7252–7261, PMLR (2019).

[9] Li, Q., He, B., and Song, D., "Model-contrastive federated learning," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 10713–10722 (2021).

[10] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K., "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360* (2016).

[11] Malekzadeh, M., Hasircioglu, B., Mital, N., Katarya, K., Ozfatura, M. E., and Gündüz, D., "Dopamine: Differentially private federated learning on medical data," *arXiv preprint arXiv:2101.11693* (2021).

[12] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A., "Automatic differentiation in pytorch," (2017).