FedSelect: Customized Selection of Parameters for Fine-Tuning during Personalized Federated Learning

Rishub Tamirisa [1 2 3] **John Won** [1 2 4] **Chengjun Lu** [1 2] **Ron Arel** [1 2 5] **Andy Zhou** [1 2]

Abstract

Recent advancements in federated learning (FL) seek to increase client-level performance by fine-tuning client parameters on local data or personalizing architectures for the local task. Existing methods for such personalization either prune a global model or fine-tune a global model on a local client distribution. However, these existing methods either personalize at the expense of retaining important global knowledge, or predetermine network layers for fine-tuning, resulting in suboptimal storage of global knowledge within client models. Enlightened by the lottery ticket hypothesis, we first introduce a hypothesis for finding optimal client subnetworks to locally fine-tune while leaving the rest of the parameters frozen. We then propose a novel FL framework, FedSelect, using this procedure that directly personalizes *both client subnetwork structure and parameters*, via the simultaneous discovery of optimal parameters for personalization and the rest of parameters for global aggregation *during training*. We show that this method achieves promising results on CIFAR-10.

1. Introduction

Federated Learning (FL) (McMahan et al., 2017) is a machine learning paradigm which utilizes multiple clients that collaborate to train models under the supervision of a central aggregator, usually referred to as the server. Unlike traditional centralized methods which required the assemblage of data at the central server, FL methods require that only parameter updates are communicated in order to coordinate the FL training process, such that any number of clients can learn from the decentralized data without direct transfer of data. This allows for maintenance of local data privacy while also providing stronger model performances than what participants could have achieved locally. Accordingly, FL has been adapted to many privacy-sensitive tasks, such as medical data classification (Sheller et al., 2020). One of the main challenges in FL is the presence of data heterogeneity, where clients' local data distributions vary significantly from one another.

This problem of data heterogeneity is most commonly addressed by *personalized federated learning (pFL)*, which adapts clients models to local distributions. Most techniques use *full model personalization*, where clients train both a personalized and a global model. However, this requires twice the computational cost of standard FL (Dinh et al., 2020; Li et al., 2021a) and is impractical in some settings. *Partial model personalization* alleviates this by splitting clients into shared and personalized parameters (Pillutla et al., 2022), where only the shared parameters are updated globally, but typically results clients that overfit to local distributions and reduced performance. Additionally, the personalized architecture needs to be manually designed before training and cannot be adapted to specific settings.

To address this, we propose *FedSelect*, where we adapt both both *architecture and parameters* for each client to its local distribution during training. Our method is based on the intuition that individual client models should choose only a necessary subset of shared parameters to encode global information for their local task, since it may not be optimal to reuse all global information from any full layer(s). We achieve this through the *Lottery Ticket Hypothesis (LTH)*, originally proposed to prune models by finding optimal subnetworks, or lottery ticket networks (LTNs) (Frankle & Carbin, 2019). However, instead of pruning the remaining parameters to zero, we reuse them as personalized parameters. We observe improved performance on CIFAR-10 compared to pruning-based LTH-FL approaches (Li et al., 2020; Mugunthan et al., 2022) and other personalized FL approaches (Liang et al., 2020; Arivazhagan et al., 2019; Collins et al., 2021; Li et al., 2021a; Oh et al., 2022) as well as reduced communication costs compared to partial model personalization.

Related Works. Partial model personalization seeks to

[1]AI@UIUC, Champaign, IL, U.S. [2]Department of Computer Science, University of Illinois at Urbana-Champaign [3]Amazon.com, Inc [4]Department of Computer Science and Engineering, POSTECH [5]Department of Psychology, University of Illinois at Urbana-Champaign. Correspondence to: Rishub Tamirisa <rishubt2@illinois.edu>.

Workshop of Federated Learning and Analytics in Practice, colocated with 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

improve the performance of client models by altering a subset of their structure or weights to better suit their local tasks. It also addresses the issue of "catastrophic forgetting" (McCloskey & Cohen, 1989), an issue in personalized FL where global information is lost when fine-tuning a client model on its local distribution from a global initialization (Kirkpatrick et al., 2017; Pillutla et al., 2022). It does this by forcefully preserving a subset of parameters, u, to serve as a fixed global representation for all clients. However, existing methods introduced for partial model personalization (Pillutla et al., 2022; Collins et al., 2021) require hand-selected partitioning of these shared and local parameters, and choose u as only the input or output layers for their experiments.

LotteryFL (Li et al., 2020) learns a shared global model via FedAvg (McMahan et al., 2017) and personalizes client models by pruning the global model via the vanilla LTH. Importantly, parameters are pruned to zero according to their magnitude after an iteration of batched stochastic gradient updates. However, due to a low final pruning percentage in LotteryFL, the lottery tickets found for each client share many of the same parameters, and lack sufficient personalization (Mugunthan et al., 2022).

2. Methods

2.1. Problem Definition

We consider a standard FL setting with N clients and one server. The set C denotes the set of client devices, where $N = |C|$. In particular, $c_k \in C$ denotes the kth client whose data distribution is given by $\mathcal{D}_{c_k} = \{x_i^k, y_i^k\}_{i=1}^{N_k}$. Let s be the number of classes assigned to the clients $c_1, \ldots, c_N$. Next, let θ denote the vector of parameters defined by the client model architecture. Then the loss of the kth client model for each data point x is $f_k(\theta_k, x)$, where θ denotes the model parameters.

While the classical FL objective shown in Equation 1 (McMahan et al., 2017) seeks to minimize loss across all clients with respect to a global parameter vector θ_G, we focus on partial model personalization.

$$\min_{\theta_G} \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{N_k} f_k(\theta_G, x_i^k) \tag{1}$$

Partial model personalization refers to the procedure in which model parameters are partitioned into shared and local parameters, denoted u and v, for averaging and local fine-tuning.

We consequently define $\theta_k = (u, v_k)$, where u denotes a set of shared global parameters, and v_k the personalized client parameters. The pFL objective following this formulation is given by:

$$\min_{u, \{v_k\}_{k=1}^{N}} \sum_{k=1}^{N} \frac{\alpha_k}{N_k} \sum_{i=1}^{N_k} f_k((u, v_k), x_i^k) \tag{2}$$

where α_k represents a constant weighting factor for aggregation of client losses.

2.2. Motivation

Prior works involving fine-tuning during both transfer learning and federated learning under distributional shift selectively fine-tune models layer-wise (Lee et al., 2023; Pillutla et al., 2022; Liang et al., 2020; Li et al., 2021b; Collins et al., 2021). In this work, we propose a novel hypothesis describing that only *parameters* that change the most during training are necessary for fine-tuning the model; the rest can be frozen as initialized parameters. Thus, drastic distributional changes in the fine-tuning task may be better accommodated by preserving pretrained knowledge *parameter-wise* rather than *layer-wise*. Following this we propose the following hypothesis:

FL Gradient-based Lottery Ticket Hypothesis. *When training a client model on its local distribution during federated learning, parameters exhibiting minimal variation are considered suitable for freezing and encoding shared knowledge, while parameters demonstrating significant fluctuation are deemed optimal for fine-tuning on local distribution and encoding personalized knowledge.*

The set of parameters selected for personalization will be trained on local data and kept locally, while the rest of the parameters that are identified as frozen will be initialized as the global parameters, then locally updated, and finally submitted to the server for federated averaging to encode shared knowledge across clients. GradLTN (Algorithm 1) describes the process by which these candidate parameters for local personalization and global updating are identified, respectively. Then, FedSelect(Algorithm 2) utilizes GradLTN's output to perform federated averaging.

2.3. Algorithms

2.3.1. GradLTN

GradLTN takes as input an initialization for the network θ_0, the number of total mask-pruning iterations L, and a mask-pruning rate r. Since statistical heterogeneity is typical across client data distributions in FL, it is a common goal to personalize client architectures or subnetworks to better adapt their local distributions. GradLTN implements this idea during the subnetwork search process, by freezing parameters that change the least, and continually fine-tuning the rest. By the end, two sets of parameters, $\theta_0 \odot \neg m_L$ and $\theta_0 \odot m_L$, are identified for averaging and fine-tuning, respectively. Although we run GradLTN for a fixed number of iterations, there are many alternative choices for the stop-

Algorithm 1 GradLTN: Gradient-based Lottery Tickets

Input: θ_0, L, r
for $i = 0$ **to** L **do**
  if $i > 0$ **then**
    $\gamma \leftarrow |\theta_i - \theta_{i-1}| \odot m_{i-1}$ *# Find new param change*
    $m_i \leftarrow$ binary mask for largest $(1-r)\%$ values in γ
    $\theta_i \leftarrow \theta_0$ *# Reinitialize model*
  else
    $m_i \leftarrow$ mask of all 1s
  end if
  for epoch $j = 1$ **to** E **do**
    $t \leftarrow 0$
    for batch $b \in B$ **do**
      # Freeze params where mask is zero
      $g_t \leftarrow \nabla_{\theta_{i,t}} l(\theta_{i,t}, b) \odot m_i$
      $\theta_{i,t+1} \odot m_i \leftarrow \theta_{i,t} \odot m_i - \eta g_t$
    end for
  end for
end for
 # Return u and v based on m_i
return $\theta_0 \odot \neg m_L, \theta_0 \odot m_L, m_L$

ping condition, i.e. setting fixed target pruning-rates and target accuracy thresholds.

For convenience, we use the Hadamard operator $\odot$ to be an indexing operator for a binary mask m, rather than an elementwise multiply operator. For example, $\theta \odot m$ assumes θ and m have the same dimensions and returns a reference to the set of parameters in θ where m is not equal to zero.

2.3.2. FedSelect

In FedSelect, the input parameters $C, \theta_G^0, K, R, L,$ and p represent clients, the first global initialization, participation rate, GradLTN iterations, and personalization rate, respectively. The key step in FedSelect is performing LocalAlt on the shared and local parameter partition identified by GradLTN. By the end of GradLTN, v_k is identified as the set of appropriate parameters for dedicated local fine-tuning via LocalAlt; u is also updated in LocalAlt and then averaged for global knowledge acquisition and retention. LocalAlt was introduced to update a defined set of shared and local parameters, u and v_k, by alternating full passes of stochastic gradient descent between the two sets of parameters (Pillutla et al., 2022). To the best of our knowledge, this is the first method to choose parameters for alternating updates in federated learning during training time.

However, averaging among the shared parameters u_k only occurs across parameters for which the corresponding mask entry in m_k is 0. This ensures that only non-LTN parameters are averaged when client models have very different masks. We store a global mask m_G to facilitate updates for clients not sampled during FL. As communication rounds progress, we hypothesize that the global knowledge stored in $\theta_G^t \odot \neg m_i^t$ is refined, and that test accuracy due to personalization will converge.

Algorithm 2 FedSelect

Input: $C = \{c_1, \ldots, c_N\}, \theta_G^0, K, R, L, p$
Server Executes:
  $k \leftarrow \max\{N \cdot K, 1\}$
  Initialize all client models $\{\theta_i^0\}_{i=1}^N$ with θ_G^0
  for each round t in $1, 2, \ldots, R$ **do**
    $S_t \leftarrow$ random sample of k clients from C
    for each client $c_k \in S_t$ **in parallel do**
      # Executed locally on client c_k
      $u_k^t, v_k^t, m_k^t \leftarrow$ GradLTN($\theta_k^{t-1}, L, 1-p$)
      ${u_k^t}^+, {v_k^t}^+ \leftarrow$ LocalAlt(u_k^t, v_k^t)
    end for
    # Averaging occurs only across clients where the mask is 1 for a given parameter's position
    $\theta_G^t \leftarrow$ Average non-LTN parameters $\{{u_k^t}^+\}_{c_k}^{S_t}$
    $m_G^t \leftarrow$ Binary OR over client masks $\bigvee_{c_k}^{S_t} m_k^t$
    for $i = 1$ **to** N **do**
      if m_i^t exists **then**
        # Distribute global params to clients' non-LTN params, located via $\neg m_i^t$
        $\theta_i^t \odot \neg m_i^t \leftarrow \theta_G^t \odot \neg m_i^t$
        $\theta_i^t \odot m_i^t \leftarrow {v_i^t}^+$
      else
        $\theta_i^t \odot m_G^t \leftarrow \theta_G^t$
      end if
    end for
  end for

An important hyperparameter of FedSelect is the personalization rate p. For different client problem difficulties (Hsieh et al., 2020), the rate of personalization p may affect the level of test accuracy achieved. A nuance of our notation is that for a given p, $(1-p) \times 100\%$ parameters are frozen during each of GradLTN's iterations. Therefore we let FedSelect (0.25) denote running FedSelect when 75% of parameters are frozen in each iteration of GradLTN. So, increasing p corresponds to fewer frozen parameters and greater personalization. Additionally, since only u_k is communicated between the server and clients, a greater p results in reduced communication costs.

A byproduct of GradLTN is that the subnetwork search process itself fine-tunes parameters in the final iterations of the algorithm, which could be valuable as an initialization for LocalAlt. Therefore, we aim to explore changing the returned values from GradLTN from θ_0 to θ_L to incorporate this idea.

3. Experiments

Models & Datasets. In this work, we consider a cross-silo setting in which the number of clients is low, but participa-

Table 1. Mean test accuracies (%) after 200 communication rounds in the full-participation, low-client setting, with 20 training samples and 100 testing samples per class for each client.

Method	$s = 2$	$s = 4$
FedAvg	33.54	52.57
LG-FedAvg	59.95	51.35
LotteryFL	73.37	60.13
Ditto	74.50	48.85
FedBABU	35.65	53.03
FedPer	30.25	51.73
FedRep	82.70	65.33
FedSelect (0.25)	84.22	65.23
FedSelect (0.50)	85.64	**65.88**
FedSelect (0.75)	**85.91**	62.91

tion is high (Liu et al., 2022). We performed all experiments using a ResNet18 (He et al., 2015) backbone pretrained on ImageNet (Deng et al., 2009). We show results for our experimental setting on non-iid samples from CIFAR-10. Each client was allocated 20 training samples and 100 testing samples per class.

Hyperparameters. We set the number of clients $|C| = N = 10$ in all experiments. The participation rate K is set to 1.0, and the number of classes per client is varied from $s = 2$ and $s = 4$. In GradLTN, we perform 5 pruning iterations, each with 5 local epochs of training. However, the personalization rate p was varied from 0.25, 0.50, and 0.75. Finally, 5 epochs of personalized training via LocalAlt are performed.

Comparisons to Prior Work. We compare our results to FedAvg (McMahan et al., 2017), LotteryFL (Li et al., 2020), FedBABU (Oh et al., 2022), FedRep (Collins et al., 2021), FedPer (Arivazhagan et al., 2019), Ditto (Li et al., 2021a), and LG-FedAvg (Liang et al., 2020). To fairly compare the performance of these methods, we fix the number of local epochs across all methods to 5. All other hyperparameters (learning rate, momentum, etc.) follow the recommended settings by the authors of the respective works.

Evaluation Metric. For all methods, the mean accuracy of the final model(s) across individual client data distributions calculated at the final communication round is reported. For FedAvg, accuracy is reported for a single global model. However, for other methods that learn personalized client models, the final average accuracy is reported by averaging individual client model accuracies.

3.1. Performance Comparison

We observe in Table 1 that FedSelect outperforms all other baselines in the CIFAR-10 in a low-client, full-participation setting. FedAvg learns a global model for all clients, resulting in reduced accuracy when confronted with increased non-IIDness. Conversely, highly personalized FL algorithms exhibit resilience to non-IIDness due to the ease of a client's local task. FedBABU and FedPer appear to suffer the same as FedAvg, despite personalizing a small set of parameters.

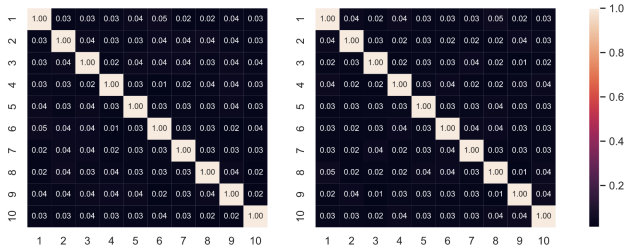

Figure 1. Intersection-over-union overlap between all pairs of client masks found by FedSelect for the final ResNet18 linear layer, for $p = 0.50$. Both the $s = 2$ masks (left) and $s = 4$ masks (right) exhibit significant diversity.

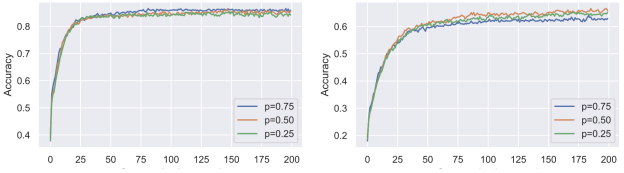

Figure 2. Average test accuracies of FedSelect on non-iid client partitions of CIFAR-10 when varying the GradLTN personalization rate p for $s = 2$ (left) and $s = 4$ (right).

Illustrated in Figure 1, the set of masks found in ResNet18's linear layer during GradLTN are significantly different from one another, indicating a high degree of personalization between clients. The resulting increased performance of FedSelect suggests that personalizing individual parameters as opposed to full layers is beneficial for FL.

3.2. Effect of Personalization Rate

In Figure 2, we find that in the $s = 2$ task, there is minimal variation in test accuracy when changing the personalization rate during FedSelect. It is possible that the process by which personalizable parameters are identified via GradLTN and LocalAlt is performed may be strong enough to perform well in the $s = 2$ task despite the low data regime. However, for $s = 4$, both high ($p = 0.75$) and low ($p = 0.25$) personalization rates perform slightly worse than $p = 0.50$. Although the rate does not directly reflect the final percentage of parameters frozen/personalized, $p = 0.5$ may represent a middle ground for which personalizing and averaging a similar number of parameters is optimal.

4. Conclusion & Future Directions

We propose FedSelect, a method for personalized federated learning that personalizes client architectures during training with the Gradient-Based Lottery Ticket Hypothesis. We demonstrate promising results on CIFAR-10, surpassing prior personalized FL and pruning-based LTH approaches in the full-participation, low-client setting. For future work, we aim to expand our method and perform extensive studies on varying the personalization rate under different settings,

and apply this technique to additional datasets, such as EMNIST (Cohen et al., 2017), Fashion MNIST (Xiao et al., 2017), and CINIC10 (He et al., 2020).

5. Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. We are thankful for the help of Chulin Xie and Wenxuan Bao for their valuable advising and support on this project. This research is part of the Delta research computing project, which is supported by the National Science Foundation (award OCI 2005572), and the State of Illinois. Delta is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. We would also like to thank Amazon, Microsoft, and NCSA for providing conference travel funding, as well as ICML for providing registration funding.

References

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers, 2019.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. Emnist: extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dinh, T. C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21394–21405, 2020.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

He, C., Li, S., So, J., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., Shen, L., Zhao, P., Kang, Y., Liu, Y., Raskar, R., Yang, Q., Annavaram, M., and Avestimehr, S. Fedml: A research library and benchmark for federated machine learning. *CoRR*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. B. The non-iid data quagmire of decentralized machine learning, 2020.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, mar 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073%2Fpnas.1611835114.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. Surgical fine-tuning improves adaptation to distribution shifts, 2023.

Li, A., Sun, J., Wang, B., Duan, L., Li, S., Chen, Y., and Li, H. Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets, 2020.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021a.

Li, X., JIANG, M., Zhang, X., Kamp, M., and Dou, Q. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021b.

Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Liu, Z., Hu, S., Wu, Z. S., and Smith, V. On privacy and personalization in cross-silo federated learning, 2022.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24: 109–165, 1989.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proc. of Int'l Conf. Artificial Intelligence and Statistics (AISTATS)*, Apr 2017.

Mugunthan, V., Lin, E., Gokul, V., Lau, C., Kagal, L., and Pieper, S. Fedltn: Federated learning for sparse and personalized lottery ticket networks. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.),

Computer Vision – ECCV 2022, pp. 69–85, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19775-8.

Oh, J., Kim, S., and Yun, S.-Y. FedBABU: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=HuaYQfggn5u.

Pillutla, K., Malik, K., Mohamed, A., Rabbat, M., Sanjabi, M., and Xiao, L. Federated learning with partial model personalization. In *International Conference on Machine Learning*, 2022.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., and Bakas, S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(12598), 2020. doi: 10.1038/s41598-020-69250-1. URL https://doi.org/10.1038/s41598-020-69250-1.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 2017.