Towards Sparsified Federated Neuroimaging Models via Weight Pruning

Dimitris Stripelis[1(✉)], Umang Gupta[1], Nikhil Dhinagar[2], Greg Ver Steeg[1], Paul M. Thompson[2], and José Luis Ambite[1]

[1] Information Sciences Institute, University of Southern California, Los Angeles, CA 90292, USA
{stripeli,umanggup,gregv,ambite}@isi.edu

[2] Imaging Genetics Center, Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA 90292, USA
{dhinagar,thompson}@ini.usc.edu

Abstract. Federated training of large deep neural networks can often be restrictive due to the increasing costs of communicating the updates with increasing model sizes. Various model pruning techniques have been designed in centralized settings to reduce inference times. Combining centralized pruning techniques with federated training seems intuitive for reducing communication costs—by pruning the model parameters right before the communication step. Moreover, such a progressive model pruning approach during training can also reduce training times/costs. To this end, we propose *FedSparsify*, which performs model pruning during federated training. In our experiments in centralized and federated settings on the brain age prediction task (estimating a person's age from their brain MRI), we demonstrate that models can be pruned up to 95% sparsity without affecting performance even in challenging federated learning environments with highly heterogeneous data distributions. One surprising benefit of model pruning is improved model privacy. We demonstrate that models with high sparsity are less susceptible to membership inference attacks, a type of privacy attack.

Keywords: Neuroimaging · Federated learning · Model pruning · Security & Privacy

1 Introduction

Federated Learning [16,18,32] enables distributed training of machine learning and deep learning models across geographically dispersed data silos. In this setting, no data ever leaves its original location, making it appealing for training

D. Stripelis and U. Gupta—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-18523-6_14.

S. Albarqouni et al. (Eds.): DeCaF 2022/FAIR 2022, LNCS 13573, pp. 141–151, 2022.
https://doi.org/10.1007/978-3-031-18523-6_14

models over private data that cannot be shared. For these reasons, Federated Learning has witnessed widespread adoption across multiple disciplines, especially in biomedical settings [3,24,26]. Federated training of neural networks involves exchanging/communicating parameters that are updated during local training on private datasets. This parameter exchange incurs high communication costs, limiting the size of neural networks that can be learned [25]. To circumvent this, model pruning techniques that have been extensively studied in centralized settings [6,9,17] for improving models' training and inference time seem a natural fit towards this direction.

In this work, we propose a federated training approach incorporating model pruning by directly extending previous work on model pruning in centralized settings [6,35]. Similar to these, we use a simple pruning approach of removing weights with the lowest magnitude. However, we consider federated learning environments with heterogeneous data distributions. The learning task is to predict brain age from T1-weighted MRI scans obtained from the UK BioBank dataset [19]. We show that with our progressive model pruning strategy, i.e., increasing the sparsity in the model with each federation round, we can learn a neural network model with less than 5% parameters of the original model while preserving most of the performance.

Even though Federated Learning avoids private data sharing, models trained using federated learning are not always private and may leak sensitive information [8,23,33]. This can often be attributed to overfitting or memorization [8,30]. Pruning parameters excessively can reduce the memorization capacity of neural networks. Inspired by this intuition, we evaluate the empirical privacy of the obtained sparsified models through membership inference attacks. We observe that pruned models at extreme degrees of sparsification (>95%) are less susceptible to membership inference attacks while maintaining learning performance. This suggests a triple win for using pruning during federated training—a) reduced communication costs, b) reduced inference costs due to small sized final models, and c) reduced privacy leakage.

Existing federated model pruning strategies focus on reducing the required communication cost during training in order to achieve specific levels of model performance [1,12]. However, in this work we aim to train highly sparsified models of similar performance to the non-sparsified counterparts while at the same time exploring the privacy gains of federated model sparsification against membership inference attacks. To the best of our knowledge, this is the first work that studies the learning performance and privacy properties of model pruning for deep learning models in the federated neuroimaging domain.

2 Neuroimaging Learning Environments

An extensive number of machine learning and deep learning approaches have been recently proposed [31] with great success [4,34] across multiple biomedical imaging tasks, such as image reconstruction, automated segmentation and predictive analytics. In this work, we evaluate such deep learning approaches for the

BrainAGE prediction task over a set of challenging neuroimaging environments in centralized and federated settings.

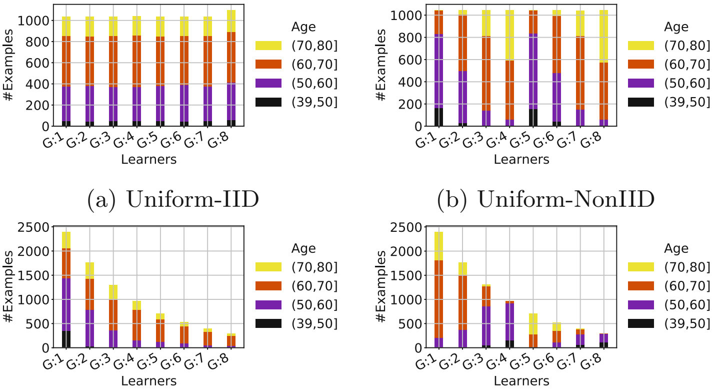
Brain Age Prediction Task. Brain age prediction involves creating a machine learning model to predict a person's chronological age from their brain MRI scan, after training the model on large amounts of data from healthy individuals. When this trained model is applied to new scans from patients and healthy controls, the age difference between each individual's true chronological age and that predicted from their MRI scan has been found to be associated with a broad range of neurological and psychiatric disorders, and with mortality [2,22]. This age prediction task is formulated as a regression task also known as the Brain Age Gap Estimation (BrainAGE). Various efficient deep learning architectures have been recently proposed based on RNNs [13,15] and CNNs [7,22] with highly accurate brain age estimations. In our work, we use a 3D-CNN model, similar to [15,27] consisting of seven blocks. The first five blocks are composed of a $3 \times 3 \times 3$ 3D convolutional layer, instance norm, a $2 \times 2 \times 2$ max-pool and ReLU activation functions. The sixth block is a $1 \times 1 \times 1$ 3D convolutional layer followed by an instance norm and ReLU activation. The final block has an average pooling layer, and a $1 \times 1 \times 1$ 3D convolutional layer. We test the performance of the model on the BrainAGE task over the UK BioBank dataset [19]. Out of the 16,356 subjects with neuroimaging in dataset, we selected 10,446 subjects with no neurological pathology and psychiatric diagnosis as defined by the ICD-10 criteria.

Centralized Environment. For centralized training, we follow the same setup as [7,15]. We consider 10,466 healthy subjects from the UKBB dataset and split them into train, test and validation sets of sizes 7,312, 2,172 and 940 respectively.

Federated Learning Environments. In our federated learning environment, we consider a centralized (star-shaped) topology [24] where a single controller orchestrates the execution of the participating learners. The controller aggregates learners' local models based on the number of training examples each model was trained on and learners train the global model on their local dataset using Vanilla SGD [18]. We refer to this federated training procedure as *FedAvg* [18].

Similar to the centralized settings, our learning task is BrainAGE prediction and the learning model is a 3D-CNN [22,27]. We partition the MRI scans of the training and validation datasets from the centralized environment across 8 learners in four federated learning environments [27,29] of heterogeneous data amounts (Uniform, Skewed) and distributions (IID, Non-IID) per learner (see Fig. 1). Uniform and Skewed refer to the cases where learners have an equal and rightly skewed number of training samples, respectively. IID and Non-IID refer to the cases where the age range of the local data distribution of the scans owned by a learner captures the global range or a subset, respectively.

Measuring Privacy via Membership Inference Attacks. To measure how much information the model leaks about the training set, we consider *Membership Inference Attack*. A Membership Inference Attack is often the preferred approach to evaluate practical privacy leakage from machine learning models [10,20].



(a) Uniform-IID

(b) Uniform-NonIID

(c) Skewed-IID

(d) Skewed-NonIID

Fig. 1. UKBB federated learning environments.

Unlike differential privacy which considers worst-case privacy leakage, membership inference attacks can be seen as evaluating average case practical privacy leakages. In particular, given a sample (a subject's brain MRI in our case), these attacks infer if the sample was used during training or not. Discovering whether the subject's MRI is in the training set can reveal the personal medical history of the subject, which is undesirable. We use the same attack setups as in [8].

In particular, for evaluating models trained in our centralized environment we use their white-box attack setup. We consider access to some actual training and unseen samples for training the attack model; this is a stronger attack setup. One can also launch attacks without accessing actual training samples by training shadow models [8,21]. We create a balanced test set of training and unseen examples, and report the accuracy of correct predictions as "attack accuracy". Lower attack accuracy is more private, and hence better.

For models trained in our federated environments, we consider one of the learners as malicious and launching attacks against other learners. In our federated environments we consider 8 learners, which translates to 56 (7×8) attacks. The learner may train attack models using their private training set and some unseen examples. We report the accuracy of correctly differentiating between other learners' training examples and unseen samples as the "attack accuracy" and report the average accuracy, as in [8]. We also report the number of successful attacks, since due to data heterogeneity not all attacks are successful. We use features derived from the predictions, labels, and gradients of the last two layers of the 3D-CNN to train the attack models.

3 Model Pruning

In this section, we discuss model pruning approach for centralized and federated environments for neuroimaging tasks. We evaluate the efficacy of the weight magnitude-based pruning approach on a 3D-CNN trained on centralized and distributed MRI scans.

Centralized Model Pruning. Neural networks can often have redundant parameters which do not affect the outcome. One of the simplest ways of identifying such parameters is by looking at the magnitude of parameters. Parameters with low absolute values do not influence the output much and thus can be safely pruned [6,35]. We use this simple approach for pruning. [35] showed that gradual parameters pruning during training is more effective than one-shot pruning at the end. Our federated pruning approach exploits this observation. However, in the centralized setting, we prune in one step at the end of 90th epoch, followed by finetuning for 10 epochs.

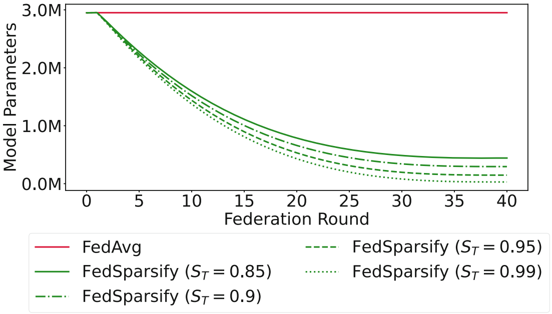


Fig. 2. Federated models number of parameters progression with (FedSparsify) and without (FedAvg) sparsification.

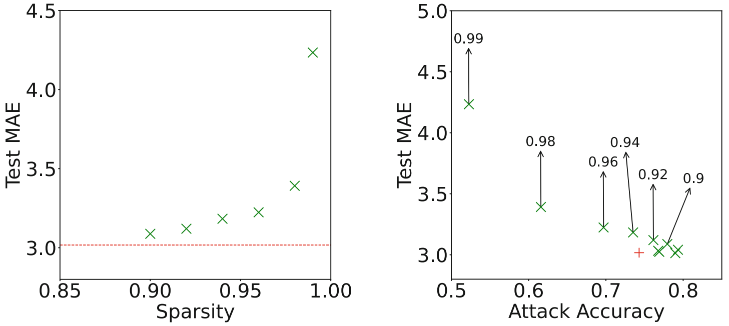
Federated Model Pruning. We develop our sparsified federated training on top of FedAvg. The global model is pruned at the controller after the controller aggregates the local model updates from the participating learners. Once the new (sparsified) global model is computed, the controller sends new global model to the learners along with the associated binary mask representing pruned and unpruned parameters. We use weight magnitude-based pruning approach [35] and remove the weights with lowest absolute values. A parameter once pruned is never resurrected. To enforce this during local training, each learner applies the binary mask at every training step (see also Algorithm 1 in Appendix). As we prune during every federation round, our pruning strategy follows a progressive schedule similar to [28,35]. The percentage of additional parameters pruned in each round follows an exponentially decreasing schedule, and the overall sparsity at round t is governed by this formula:

$$s_t = S_T + (S_0 - S_T)\left(1 - \frac{F\lfloor t/F \rfloor - t_0}{T - t_0}\right)^n \tag{1}$$

Here T is total number of federation rounds, S_0 and S_T are the initial and desired final sparsity, F is frequency of sparsification, and t_0 is the initial sparsification round. The exponent n controls the exponential sparsification rate. We refer to this pruning strategy as *FedSparsify*. In our experimental evaluation, we explore different final sparsities, i.e., $S_T = \{85\%, 90\%, 95\%, 99\%\}$. Throughout our experiments, we set the rate of sparsification n to 3, we prune the global model at every federation round, i.e., $F = 1$, for a total number of 40 federation rounds, $T = 40$, and we start the sparsification schedule at federation round 1, $t_0 = 1$. Figure 2 presents the progression of global model parameters of this sparsification schedule over the course of 40 federation rounds.

4 Results

We train the 3D-CNN model[1] for the brain age prediction task in different learning setups. We perform one-shot pruning in the centralized setup to achieve different sparsity levels. For the federated learning setup, we vary S_T, the final sparsity level in Eq. 1 and prune progressively before communicating updated weights to the learners (see Algorithm 1). In all environments the model is trained using Vanilla SGD with a batch size of 1 and learning rate of $1e^{-5}$. During federated training learners train the global model locally for 4 epochs in between federation rounds. All experiments were run on a dedicated GPU server equipped with 4 Quadro RTX 6000/8000 graphics cards of 50 GB RAM each, 31 Intel(R) Xeon(R) Gold 5217 CPU @ 3.00 GHz, and 251 GB DDR4 RAM.



(a) (b)

Fig. 3. Centralized BrainAGE model performance at different sparsity levels (left plot) and model vulnerability to membership inference attacks with respect to model performance (right plot).

Model Pruning Does Not Hurt Performance. We first study model performance at different sparsity levels by evaluating the models on a held-out test set. These results are summarized in Fig. 3a for centralized training. Even

[1] https://github.com/dstripelis/FedSparsify.

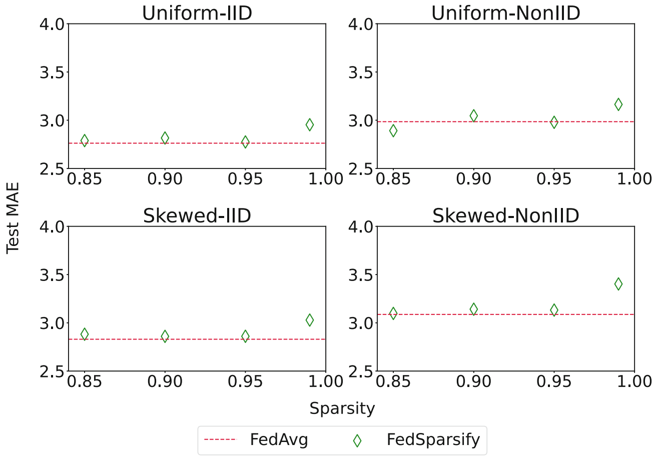


Fig. 4. Federated BrainAGE models learning performance at different degrees of sparsification across all four federated learning environments. Dashed line represents performance of non-sparsified model.

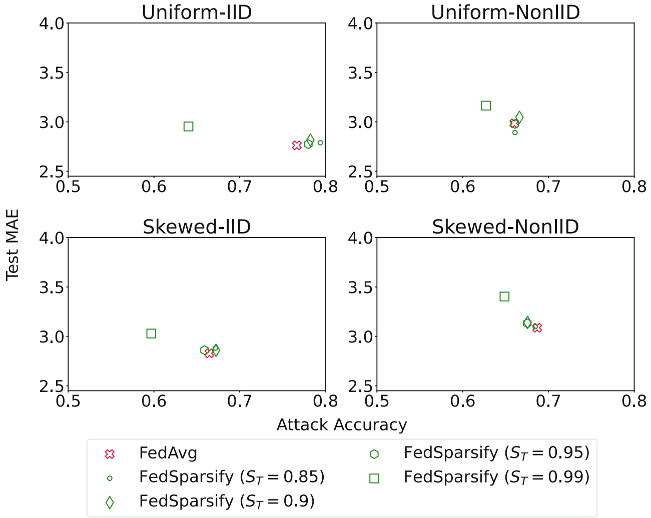


Fig. 5. Federated BrainAGE models vulnerability to membership inference attacks with respect to learning performance across all federated environments.

through the one-step pruning approach, we observe that most of the model performance is preserved when 90% of the parameters are removed. This validates the applicability of weight magnitude-based pruning for deep learning models on neuroimaging tasks. We apply our proposed progressive pruning procedure for federated training at different final sparsity levels across four different environments. The results are summarized in Fig. 4. In all cases, model performance is

Table 1. Federated models comparison in the *Skewed-IID* environment.

Sparsity	Params	Size (MBs)	Comm. (MM)	Test MAE	MIA (Success)	Throughput
0.0	2,950,401	10.85	1888	2.879	0.66 (50)	64.31
0.85	442,561	2.09	714	2.881	0.671 (52)	69.06
0.9	295,041	1.43	645	2.859	0.672 (51)	71.28
0.95	147,521	0.73	576	2.861	0.659 (54)	78.27
0.99	29,505	0.16	521	3.024	0.596 (47)	128.55

not affected at 95% sparsity level and performs the same as the FedAvg model, which is trained without pruning. Even when only 1% of the parameters are preserved, i.e., 99% sparsity, the model performance degrades slightly. Table 1 provides a quantitative comparison of the total number of parameters and memory/disk size of the final model, the cumulative communication cost in terms of the total number of parameters exchanged during training[2], and the model's learning performance. Our pruning schedule can learn a highly sparsified federated learning model with 3 to 3.5 times lower communication cost than its unpruned counterpart (cf. 521 million to 1888 million parameters). Moreover, the reduced number of the final model parameters also leads to reduced model space/memory footprint, with the sparsified models at 95% and 99% sparsification being 67 times smaller than the original model. Following previous work [14] on model efficiency evaluation[3], we benchmark the inference time for sparse and non-sparse models by recording the total number of processing items per second (i.e., Throughput - items/sec) that each model can perform. Specifically, we take the final model learned with (FedSparsify) and without sparsfication (FedAvg) and stress test its inference time by allocating a total execution time of 60 s with a warmup period of 10 s. As we show in Table 1, as sparsification increases model throughput increases too, leading to improved inference efficiency especially at 99% sparsity.

Excessive Model Pruning May Reduce Privacy Vulnerability. Intuitively, pruning can reduce the ability of a neural network to memorize training data and thus reduce privacy vulnerability. To this end, we evaluate pruned models for privacy leakage using membership inference attacks (Fig. 3b and Fig. 5). We find that at extreme sparsity levels (>95% for centralized settings and 99% for federated setting) the attack accuracy reduces suggesting that these models are less vulnerable to privacy leakage compared to non-sparsified models. Compared to the non-sparsified model, the sparsified models are 10% to 20% less vulnerable in case Skewed IID and Uniform IID environments, respectively, and 5% for the Non-IID environments.

[2] Communication cost is computed as $\sum_t^T 2N_Z^t L$. T represents the total number of federation rounds, N_Z^t the non-zero model parameters at round t and L the number of participating learners. Factor 2 accounts for the model parameters sent from the controller to the learners and from the learners to the controller within a round.

[3] https://github.com/neuralmagic/deepsparse.

5 Discussion

We investigated model pruning for deep learning models in the neuroimaging domain through the BrainAGE prediction task in both centralized and federated learning environments. We demonstrated that sparsified models are equally performant as their non-sparsified counterparts even at extreme sparsity levels across all investigated environments. We also evaluated the effectiveness of sparsified models in improving model resiliency against membership inference attacks. We discovered that highly sparsified models could reduce vulnerability to this privacy attack. The vulnerability to membership inference attack is related to the mutual information between the training dataset and activations [11] or model parameters [5]. These results could provide a plausible theoretical explanation as to why pruning reduces the information about the training dataset in neural network weights compared to weights obtained by training without pruning. In the future, we plan to analyze the relation between model sparsification and model privacy and provide a theoretical framework to understand the connection between them better. We also plan to improve model privacy by introducing notions of stochasticity while applying model weight pruning.

References

1. Bibikar, S., Vikalo, H., Wang, Z., Chen, X.: Federated dynamic sparse training: computing less, communicating less, yet learning better (2021)
2. Cole, J.H., Leech, R., Sharp, D.J., Alzheimer's Disease Neuroimaging Initiative: Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann. Neurol. **77**(4), 571–581 (2015)
3. Dayan, I., et al.: Federated learning for predicting clinical outcomes in patients with Covid-19. Nat. Med. **27**(10), 1735–1743 (2021)
4. Ezzati, A., et al.: Predictive value of ATN biomarker profiles in estimating disease progression in Alzheimer's disease dementia. Alzheimer's & Dementia **17**(11), 1855–1867 (2021)
5. Farokhi, F., Kaafar, M.A.: Modelling and quantifying membership information leakage in machine learning. arXiv preprint arXiv:2001.10648 (2020)
6. Frankle, J., Carbin, M.: The lottery ticket hypothesis: finding sparse, trainable neural networks. In: International Conference on Learning Representations (2018)
7. Gupta, U., Lam, P.K., Ver Steeg, G., Thompson, P.M.: Improved brain age estimation with slice-based set networks. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 840–844. IEEE (2021)
8. Gupta, U., Stripelis, D., Lam, P.K., Thompson, P., Ambite, J.L., Ver Steeg, G.: Membership inference attacks on deep regression models for neuroimaging. In: Medical Imaging with Deep Learning, pp. 228–251. PMLR (2021)
9. Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., Peste, A.: Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. J. Mach. Learn. Res. **22**(241), 1–124 (2021)
10. Jayaraman, B., Wang, L., Evans, D., Gu, Q.: Revisiting membership inference under realistic assumptions. arXiv preprint arXiv:2005.10881 (2020)
11. Jha, S.K., et al.: An extension of Fano's inequality for characterizing model susceptibility to membership inference attacks. arXiv preprint arXiv:2009.08097 (2020)

12. Jiang, Y., et al.: Model pruning enables efficient federated learning on edge devices. IEEE Trans. Neural Netw. Learn. Syst. (2022)
13. Jónsson, B.A., et al.: Brain age prediction using deep learning uncovers associated sequence variants. Nat. Commun. **10**(1), 1–10 (2019)
14. Kurtz, M., et al.: Inducing and exploiting activation sparsity for fast inference on deep neural networks. In: International Conference on Machine Learning, pp. 5533–5543. PMLR (2020)
15. Lam, P.K., et al.: Accurate brain age prediction using recurrent slice-based networks. In: 16th International Symposium on Medical Information Processing and Analysis, vol. 11583, p. 1158303. International Society for Optics and Photonics (2020)
16. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Signal Process. Mag. **37**(3), 50–60 (2020)
17. Liu, Z., Sun, M., Zhou, T., Huang, G., Darrell, T.: Rethinking the value of network pruning. In: International Conference on Learning Representations (2018)
18. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
19. Miller, K.L., et al.: Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nat. Neurosci. **19**(11), 1523–1536 (2016)
20. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: IEEE Symposium on Security and Privacy (SP) (2019)
21. Nasr, M., Shokri, R., Houmansadr, A.: Machine learning with membership privacy using adversarial regularization. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 634–646 (2018)
22. Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M.: Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. **68**, 101871 (2021)
23. Pustozerova, A., Mayer, R.: Information leaks in federated learning. In: Proceedings of the Network and Distributed System Security Symposium, vol. 10 (2020)
24. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**(1), 1–7 (2020)
25. Ro, J.H., et al.: Scaling language model size in cross-device federated learning. In: ACL Workshop on Federated Learning for Natural Language Processing (2022)
26. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**(1), 1–12 (2020)
27. Stripelis, D., Ambite, J.L., Lam, P., Thompson, P.: Scaling neuroscience research using federated learning. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1191–1195. IEEE (2021)
28. Stripelis, D., Gupta, U., Steeg, G.V., Ambite, J.L.: Federated progressive sparsification (purge, merge, tune)+. arXiv preprint arXiv:2204.12430 (2022)
29. Stripelis, D., Thompson, P.M., Ambite, J.L.: Semi-synchronous federated learning for energy-efficient training and accelerated convergence in cross-silo settings. ACM Trans. Intell. Syst. Technol. (TIST) (2022)
30. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Towards demystifying membership inference attacks. arXiv preprint arXiv:1807.09173 (2018)
31. Wainberg, M., Merico, D., Delong, A., Frey, B.J.: Deep learning in biomedicine. Nat. Biotechnol. **36**(9), 829–838 (2018)
32. Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., Yu, H.: Federated learning. Synthesis Lectures Artif. Intell. Mach. Learn. **13**(3), 1–207 (2019)

33. Zari, O., Xu, C., Neglia, G.: Efficient passive membership inference attack in federated learning. In: NeurIPS PriML Workshop (2021)
34. Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S.: Image reconstruction by domain-transform manifold learning. Nature **555**(7697), 487–492 (2018)
35. Zhu, M., Gupta, S.: To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878 (2017)