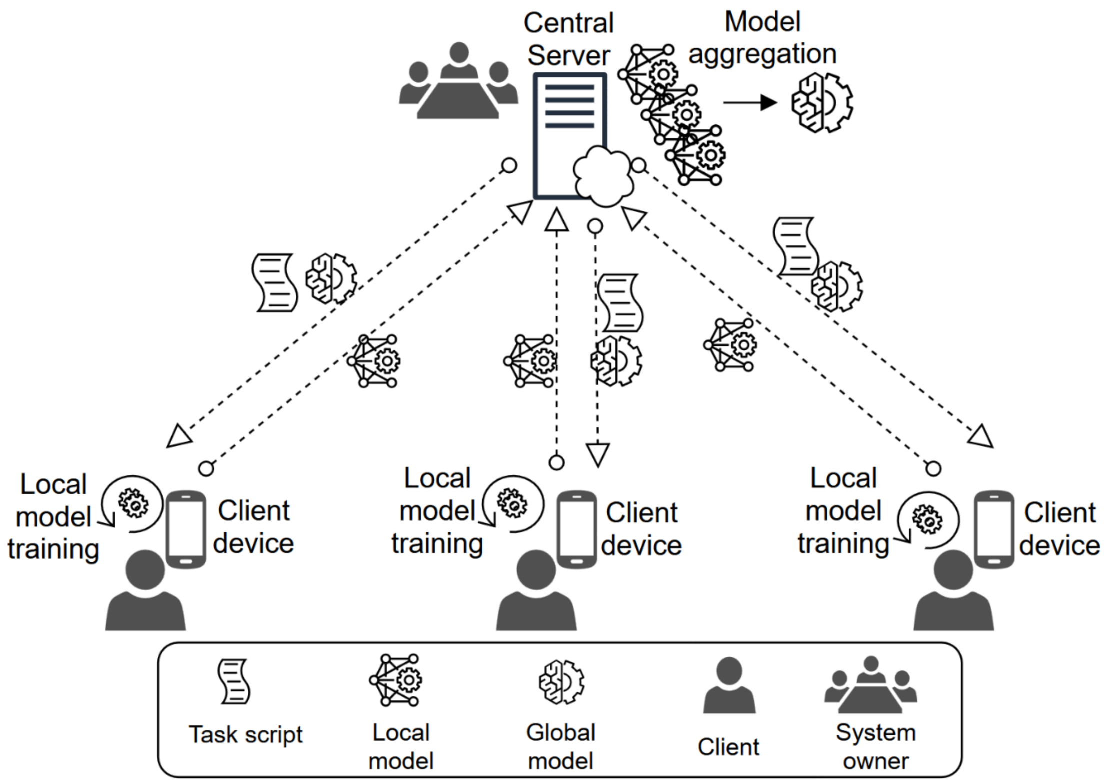
FSL-SAGE: Accelerating Federated Split Learning via Smashed Activation Gradient Estimation

Srijith Nair[1] Michael Lin[1] Peizhong Ju[2] Amirezza Talebi[1] Elizabeth Bentley[3] Jia Liu[1]

[1]The Ohio State University [2]University of Kentucky [3]Air Force Research Laboratory

Motivation

- Distributed training of large DNN on commodity devices containing private datasets.
- Federated Learning (FL) trains a model on several client datasets without sharing data.
- Model is trained in parallel on clients and periodically aggregated at server.
- Fast, but assumes clients have enough resources to store and train large models.
- Impractical for today's LLMs and foundation models.



Prior State of the Art

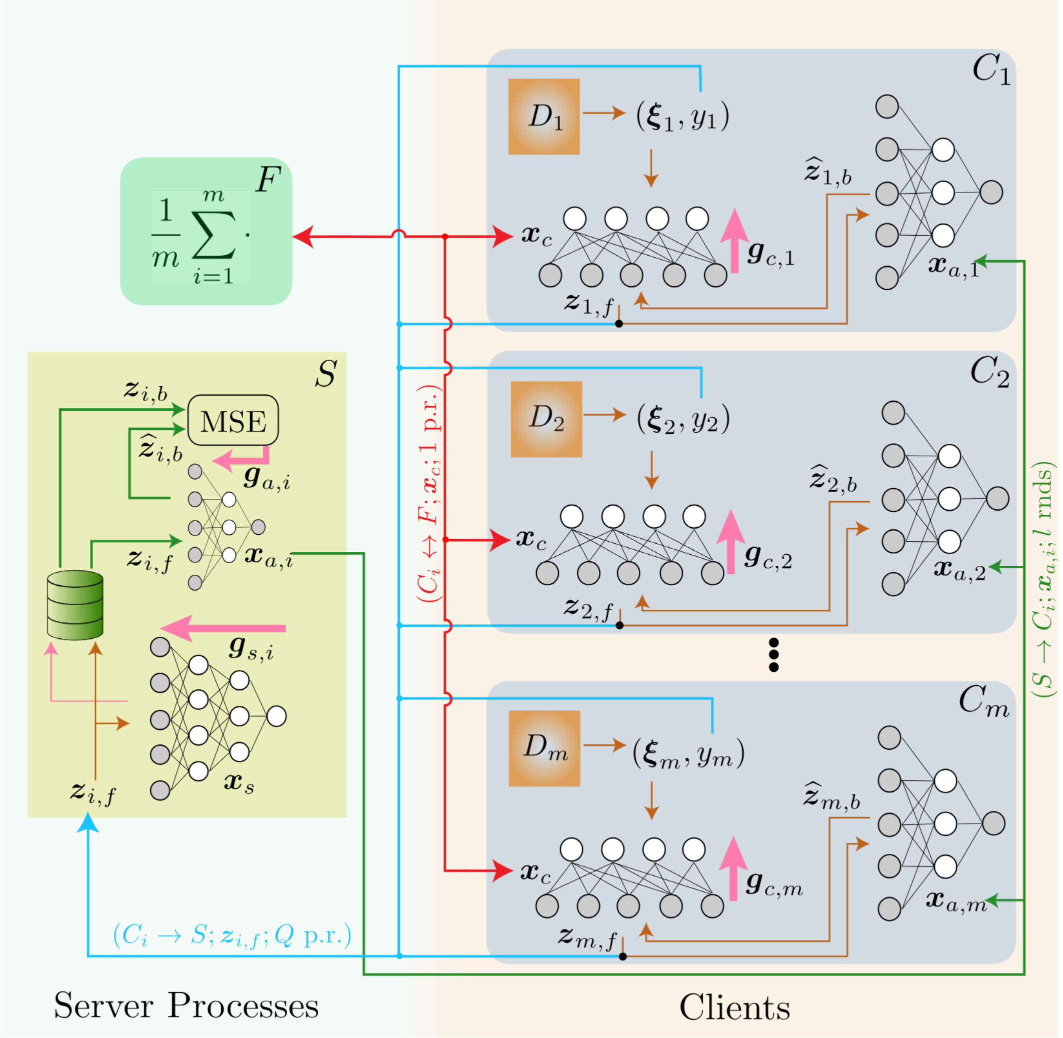
- **Split Learning (SL)**: split model between client and server; sequentially process clients in a round-robin manner
  - [Vepakomma et. al 2018, Gupta & Raskar 2018]
  - **Limitation:** low speed due to highly sequential processing; high communication load between clients and server

- **Split Federated Learning (FSL)**: two variants of algorithms: 1) SFLv1 trains one copy of server-side model for each client 2) SFLv2 sequentially updates single copy of server-side model
  - [Thapa et. al 2022]
  - **Limitation:** high server memory usage; same communication load as split learning

- **FSL with auxiliary models**: in SL setup, use local loss functions at client to approximate server-side model
  - [Han et. al 2021, Mu & Shen 2023]
  - **Limitation:** lower accuracy compared to SL; lack of server feedback when training auxiliary models; lack theoretical convergence guarantees on global model

FSL: Proposed Solution

- **FSL-SAGE:** Smashed Activation Gradient Estimation
- Auxiliary Models (AM) are explicitly trained to mimic the server-side model
- FSL-SAGE enjoys a finite-time convergence guarantee; first of its kind.
- Higher accuracy, robustness and communication efficiency compared to previous state-of-the-art



Convergence of FSL-SAGE

Theorem: Convergence Rate

Under above assumptions and step-sizes (η, η_L) *for* T *rounds, the iterates in FSL-SAGE satisfy:*

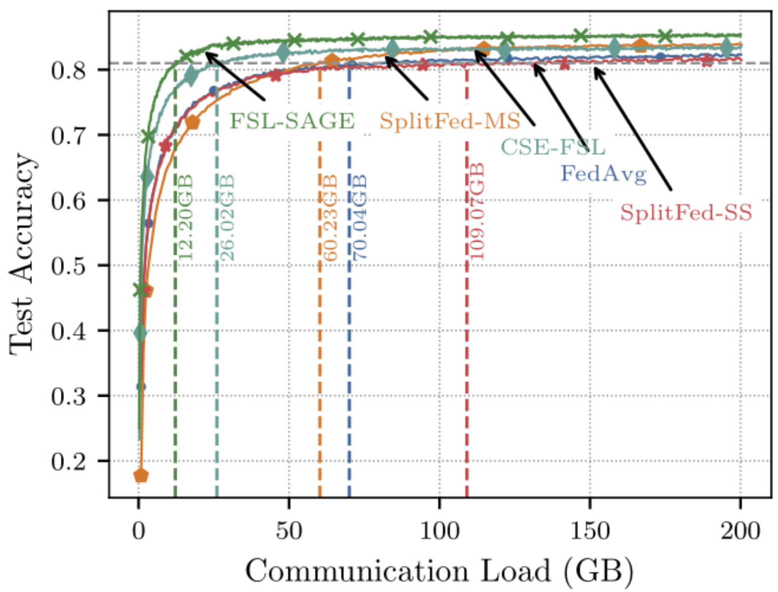
$$\min_{n\in\{1,\ldots,\lfloor T/l\rfloor\}} \mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}^{nl-1})\right\|^2\right] \le \frac{f(\boldsymbol{x}_0)-f^*}{c\min\{\boldsymbol{\eta_L}, m\boldsymbol{\eta}\}Q\boldsymbol{T}} + \frac{3CK\boldsymbol{\eta_L}}{2Q\min\{\boldsymbol{\eta_L}, m\boldsymbol{\eta}\}\sqrt{\boldsymbol{T}}} + \frac{\Phi(\boldsymbol{\eta_L}, \boldsymbol{\eta})}{\boldsymbol{T}} + \frac{3K\eta_L L_f^2}{2cQ\min\{\boldsymbol{\eta_L}, m\boldsymbol{\eta}\}}\frac{\mathbf{1}}{\boldsymbol{T}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{T}}\varepsilon_\star^t$$

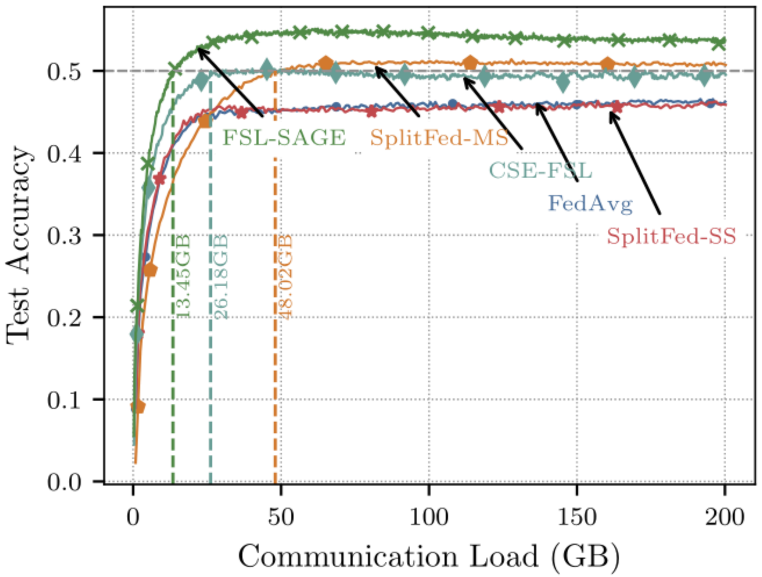
where $C>0$ *and* $c>0$ *are some constants, and* $\varepsilon_\star^t := \frac{1}{m}\sum_{i=1}^m \mathcal{L}_i(\boldsymbol{x}_{a,i}^{t\star}, \boldsymbol{x}^t)$.

- With suitable step size choices (η, η_L), convergence rate is $\mathcal{O}(1/\sqrt{T})$ for T rounds
- Last term reveals the role on the *learnability of the auxiliary model*

Experimental Results I

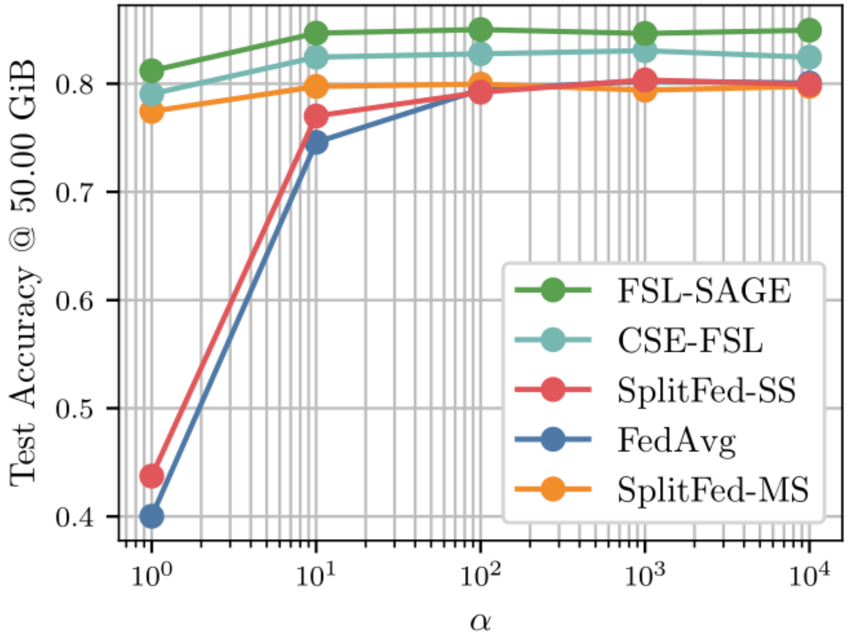
- Accuracy vs. Communication Load performance of FSL-SAGE with baselines
- Performance on image classification task: CIFAR-10 (above) and CIFAR-100 (below)
- FSL-SAGE outperforms all baselines in terms of final accuracy
- Achieves comparable accuracy with $\approx 2\times$ the communication load





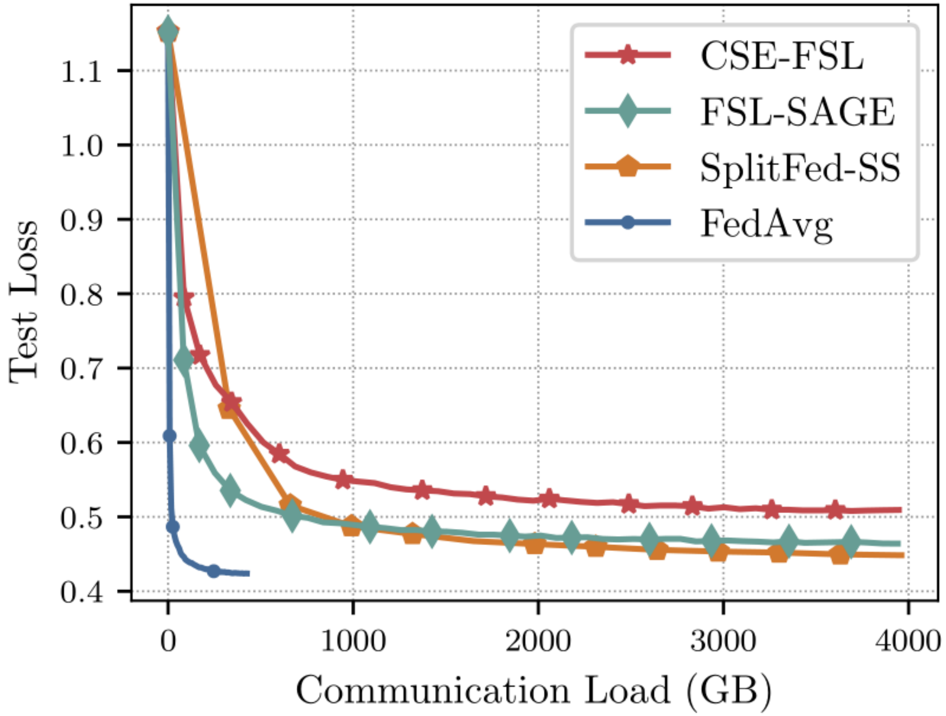
FSL: Experimental Results II

- Accuracy vs. heterogeneity in client data
- Heterogeneity measured in terms of α: Lower α implies higher heterogeneity
- FSL-SAGE is most robust to client data heterogeneity among all methods



Experimental Results III

- Preliminary results on LLM finetuning use-case: Test loss vs. communication load
- GPT-2 medium model fine-tuned to perform text completion on E2E dataset
- FSL-SAGE performs comparably to SplitFed-SS (demonstrates convergence accuracy)
- Main contender CSE-FSL is not as accurate



Thank You!