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

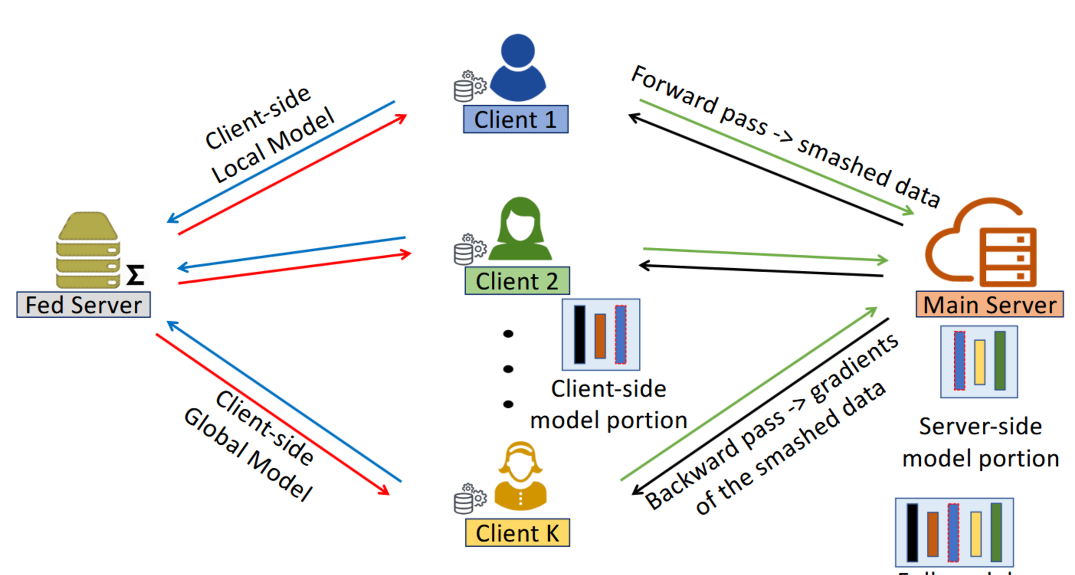
Srijith Nair[1] Michael Lin[1] Peizhong Ju[2] Amirezza Talebi[1] Elizabeth Bentley[3] Jia Liu[1]

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

Summary

- **Problem:** Traditional Federated Learning (FL) is fast but assumes that clients can store and train large-scale models; Recent works have tried to combine split learning (SL) with FL (jointly called FSL), but are communication inefficient and/or learn a suboptimal joint model.
- **Solution:** We propose to use *auxiliary models* as estimators of the server-side model in FSL; our proposed algorithm, FSL-SAGE, *provably converges* at a $\mathcal{O}\left(1/\sqrt{T}\right)$ rate for T rounds, and we save communication costs by more than 50%.

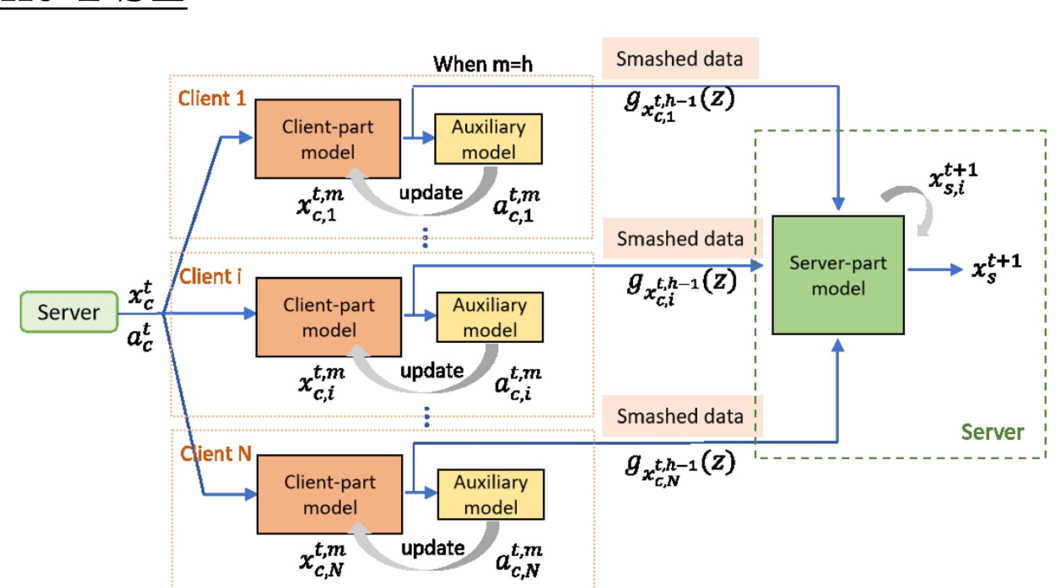
Federated Split Learning (FSL)



Picture credit (Thapa et al. 2022)

- Model split into client-side (CSM) and server-side (SSM)
- CSMs train in parallel → send data to SSM → wait for SSM to sequentially process each request
- Drawback: *Slow and communication inefficient*

Efficient FSL



Picture credit (Mu & Shen 2022)

- Use *local auxiliary models* to train CSMs in parallel
- Drawback: Sub-optimal due to *lack of SSM feedback*

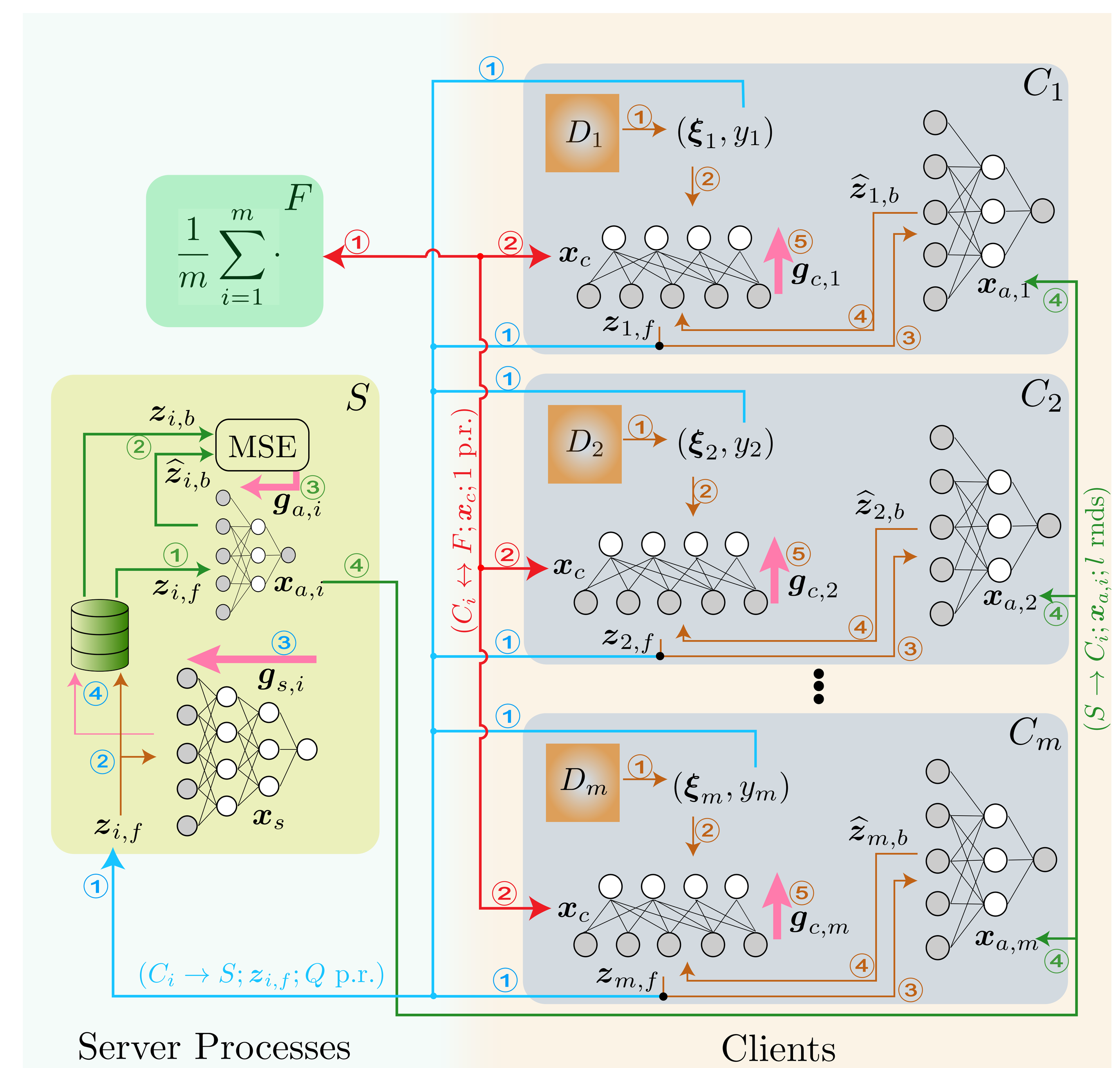
Smashed Activation Gradient Estimation (SAGE)

Idea: *Use auxiliary models (AMs) as explicit estimators of gradients returned by the SSM.*

1. Train CSMs in parallel via AMs: every iter
2. Update SSM at S: every K/Q iters
3. Aggregate CSMs at F: every K iters := 1 round.
4. *Align* AMs at S: every l rounds

Alignment:

$$\min_{\boldsymbol{x}_a} \frac{1}{2}\sum_{j=1}^{Qt}\left\|\widehat{\boldsymbol{z}}_{i,b}(\boldsymbol{x}_a)-\boldsymbol{z}^j_{i,b}\right\|_2^2$$



Finite-Time Convergence Guarantee

- First convergence guarantee on the joint model for AM based FSL!
- Assumptions: (1) loss function smoothness, (2) bounded heterogeneity of gradients (standard in FL); and, (3) *In-Expectation Learnability* of AMs.

Definition (In-Expectation Learnability): The function $\boldsymbol{f}(\cdot)$ is *in-expectation learnable* by class $\mathcal{G}$ parametrized by $\boldsymbol{\theta}$ iff $\forall\,\epsilon>0, \exists\, r_{\mathcal{G}}(\epsilon)$ such that, for $r \geq r_{\mathcal{G}}(\epsilon)$ training samples ($\mathbb{E}[\cdot]$ taken over $\widehat{\boldsymbol{\theta}}_r$ & $\boldsymbol{x}$):

$$\mathbb{E}\left[\left\|\boldsymbol{g}(\widehat{\boldsymbol{\theta}}_r;\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x})\right\|^2\right] \leq \min_{\boldsymbol{\theta}}\mathbb{E}\left[\left\|\boldsymbol{g}(\boldsymbol{\theta};\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x})\right\|^2\right]+\epsilon.$$

Theorem: *At the T^{th} round for step-sizes (η,η_L), the joint model $\boldsymbol{x}$ satisfies:*

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

where $C>0$ and $c>0$ are constants, and $\varepsilon^t_\star$ is the minimum estimation error of AM at the t^{th} round.

Corollary: For the step-size choices $\eta_L = \mathcal{O}(1/\sqrt{T})$ and $\eta = \mathcal{O}(1/(m\sqrt{T}))$, the non-lazy FSL-SAGE with PAC-learnable auxiliary models achieves a finite-time convergence rate of $\mathcal{O}(1/\sqrt{T})+\mathcal{O}(1/T)\sum_{t=1}^{T}\varepsilon^t_\star$.

Takeaways:

- *Convergence rate is equal to that of FedAvg (FL)*
- *Last term in Corollary is irreducible error; depends on the architecture of the AM*

Experimental Results

1) Image Classification Results

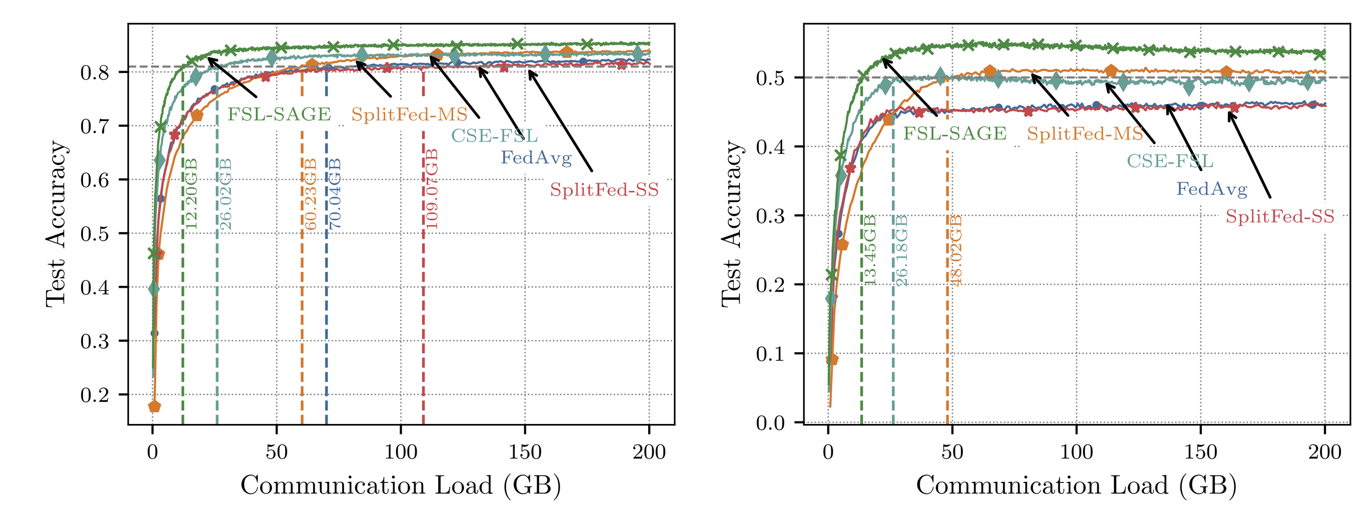


Figure 1 (Efficiency): Accuracy vs. comm load for ResNet-18 on CIFAR-10 (L) & CIFAR-100 (R); *(Takeaway: FSL-SAGE saves comm load by more than 50%)*

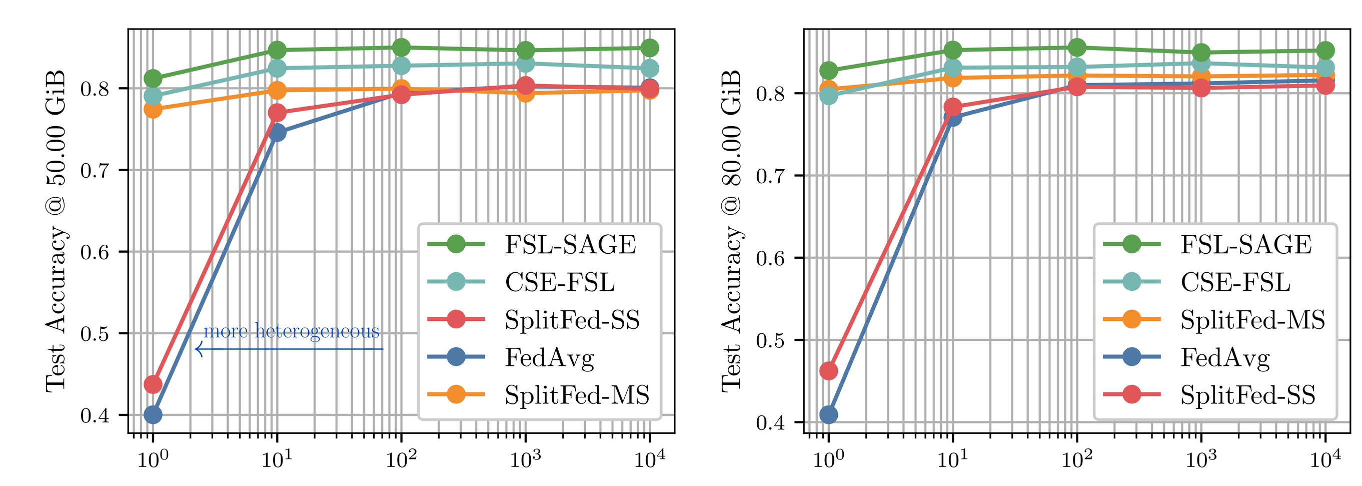


Figure 2 (Robustness): Best accuracy up to comm budget vs. data heterogeneity on CIFAR-10; *(Takeaway: FSL-SAGE is very robust to data heterogeneity)*

2) LoRA Fine-Tuning:

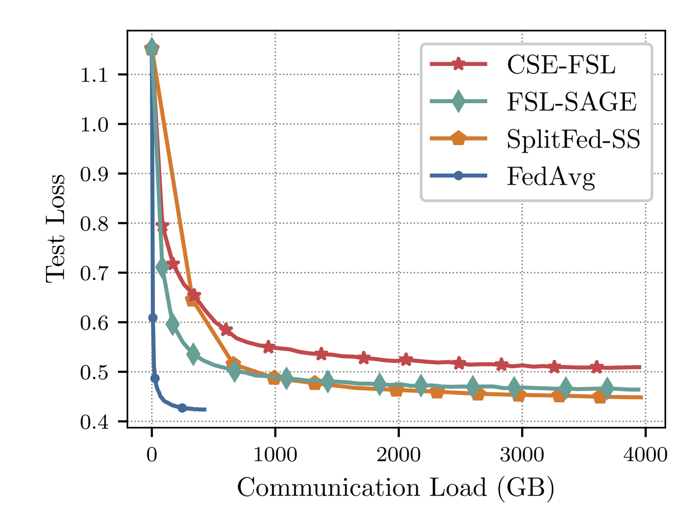


Figure 3 (LLM): Test loss vs. comm load for GPT2-m on WebNLG E2E dataset using LoRA. *(Takeaway: FSL-SAGE attains loss comparable to SL)*

Acknowledgments

This work is supported in part by ONR grant N00014-24-1-2729; NSF grants CAREER 2110259, 2112471, and 2324052; DARPA YFA D24AP00265 and DARPA HR0011-25-2-0019.

Correspondence to:

- Srijith Nair (nair.203@osu.edu)
- Jia Liu (liu@ece.osu.edu)

Arxiv paper

Source code