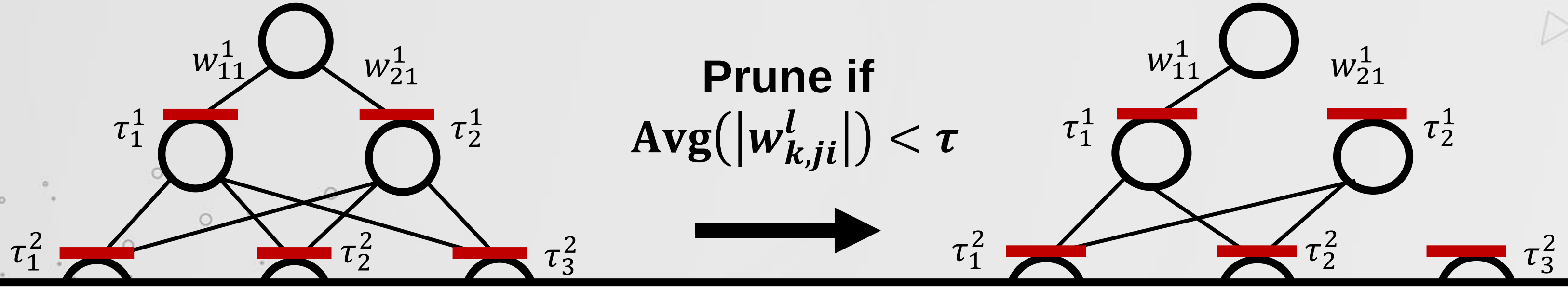
SpaFL: Communication-efficient FL with Sparse Models with Low Computational Overhead

Minsu Kim, Walid Saad, Merouane Debbah, and Choong Seon Hong

Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Dec. 2024

SpaFL Framework for Learning Sparse Structures

- **What is SpaFL?**
  - It is for learning structured sparsity across clients with limited computing and communication resources
  - Can clients collaborate to learn optimal sparse structures without sending parameters?
- **How does SpaFL make structured sparsity?**
  - We first define a learnable threshold τ for each neuron/filter
    → can be applied to MLP, CNN, and Attention layers
  - Prune entire neuron/filter if its connected average weights is smaller than the threshold



Thresholds represent how important the connected parameters are

Problem Formulation

NEWS@VT

- **How can clients learn the optimal sparse structures with thresholds $\boldsymbol{\tau}$?**

$$\min_{\boldsymbol{\tau}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_N} \sum_{k=1}^{N} F_k(\tilde{\boldsymbol{w}}_k, \boldsymbol{\tau}),$$

$$\text{s.t.} \quad F_k(\tilde{\boldsymbol{w}}_k, \boldsymbol{\tau}) = \frac{1}{D_k} \sum_{i=1}^{D_k} \mathcal{L}(\boldsymbol{w}_k \odot \boldsymbol{p}_k(\boldsymbol{\tau}); \{\boldsymbol{x}_i, y_i\}),$$

$$\boldsymbol{p}_k(\boldsymbol{\tau}) = \{\boldsymbol{p}_k^l(\boldsymbol{\tau})\}_{l=1}^{L} = \{S(\bar{\boldsymbol{w}}_k^l - \boldsymbol{\tau}^l)\}_{l=1}^{L}$$

Global thresholds

Model parameters

Loss function

Input

Label

Binary masks

# of layers

Step function

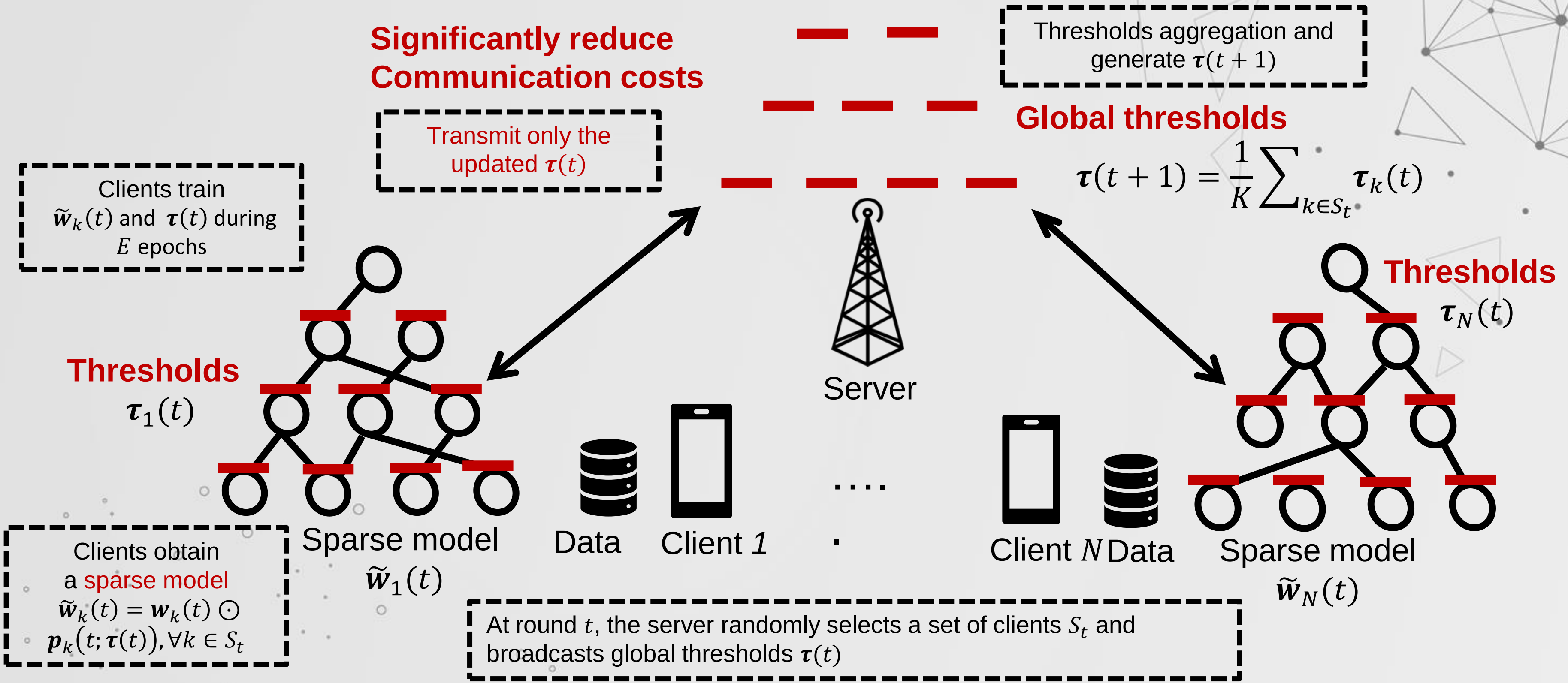
- **Does it really work?**
  - We only trained threshold τ while freezing model parameters w

	FMNIST	CIFAR-10	CIFAR-100
Only trained τ	65.62 ± 5.3	60.94 ± 3.4	24.80 ± 1.1
Initialization	10	10	1

Learning spare structures can improve the performance

SpaFL Flow

- SpaFL only communicates updated thresholds τ between the server and clients



SpaFL Generalization Analysis

NEWS@VT

- **SpaFL only communicates updated thresholds τ between the server and clients**

Theorem 1. *For the loss function $||\mathcal{L}||_\infty \leq 1$, the training data size $D \geq \frac{2}{\epsilon'^2} \ln\left(\frac{16}{\exp(-\epsilon'\delta')}\right)$ and the total number of communication rounds T, we have*

$$\mathbb{P}\left[\left|\hat{\mathcal{R}}(\mathcal{A}(\mathcal{D})) - \mathcal{R}(\mathcal{A}(\mathcal{D}))\right| < 9\epsilon'\right] > 1 - \frac{\exp(-\epsilon')\delta'}{\epsilon'} \ln\frac{2}{\epsilon'}, \quad (14)$$

where $\epsilon' = \sqrt{2T\log\frac{1}{\delta}\tilde{\epsilon}^2 + T\tilde{\epsilon}\frac{\exp(\tilde{\epsilon})-1}{\exp(\tilde{\epsilon})+1}}$,

Generalization error

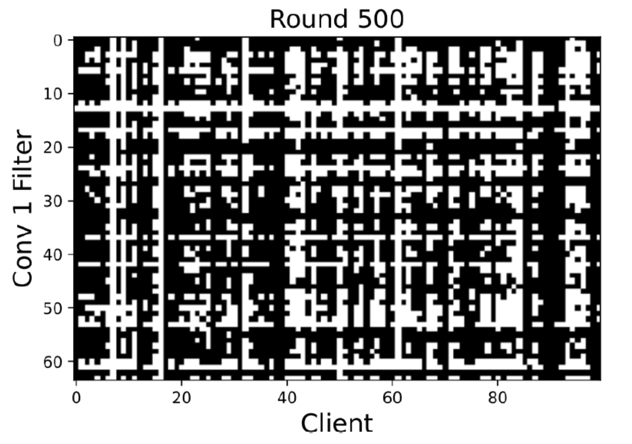
Decreasing function of model density

- **As models become more sparse, the generalization error bound becomes tighter**
- **SpaFL can improve the generalization error by learning optimal sparse structures by communicating thresholds τ**

Simulation Results

- **Performance comparison with other SOTA baselines**

	FMNIST			CIFAR10			CIFAR100		
Algorithms	Acc	Comm (Gbit)	FLOPs (e+11)	Acc	Comm (Gbit)	FLOPs (e+13)	Acc	Comm (Gbit)	FLOPs (e+14)
SpaFL	89.21±0.25	**0.1856**	**2.3779**	**69.75±2.81**	**0.4537**	**1.4974**	**40.80±0.54**	**4.6080**	**1.2894**
FedAvg	88.73±0.21	133.8	10.345	61.33±0.15	258.36	12.382	35.51±0.10	10712	8.7289
FedPM	63.27± 1.65	66.554	5.8901	52.05± 0.06	133.19	7.0013	28.56 ± 0.15	5506.1	5.423
HeteroFL	85.97±0.20	68.88	5.1621	66.83±1.15	129.178	6.1908	37.82±0.15	5356.4	4.3634
Fjord	89.08±0.17	64.21	5.1311	66.38±2.01	128.638	6.1428	39.13±0.22	5251.4	4.1274
FedSpa	**89.30±0.20**	55.256	5.2510	67.03±0.63	129.31	4.2978	36.32±0.35	5342.2	9.275
FedP3	89.12±0.14	41.327	5.8923	67.54±0.52	67.345	6.8625	37.73±0.42	2682.6	4.9384
Local	84.31±0.20	0	3.7982	57.06±1.30	0	1.9373	33.77±1.87	0	1.5384



SpaFL outperforms other baselines with less computing and communication resources

Visualization of a learned conv layer on CIFAR10

NEWS@VT

Thank you!

Question: msukim@vt.edu