Adaptive Channel Sparsity for Federated Learning under System Heterogeneity

Paper tag: THU-AM-376

Dongping Liao[1], Xitong Gao[2], Yiren Zhao[3], Chengzhong Xu[1]

[1]University of Macau, [2]Shenzhen Institutes of Advanced Technology, [3]Imperial College London

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU
DEPARTMENT OF
COMPUTER AND INFORMATION SCIENCE

Quick Preview of Our Work

Flado: Adaptive Channel Sparsity for Federated Learning under System Heterogeneity

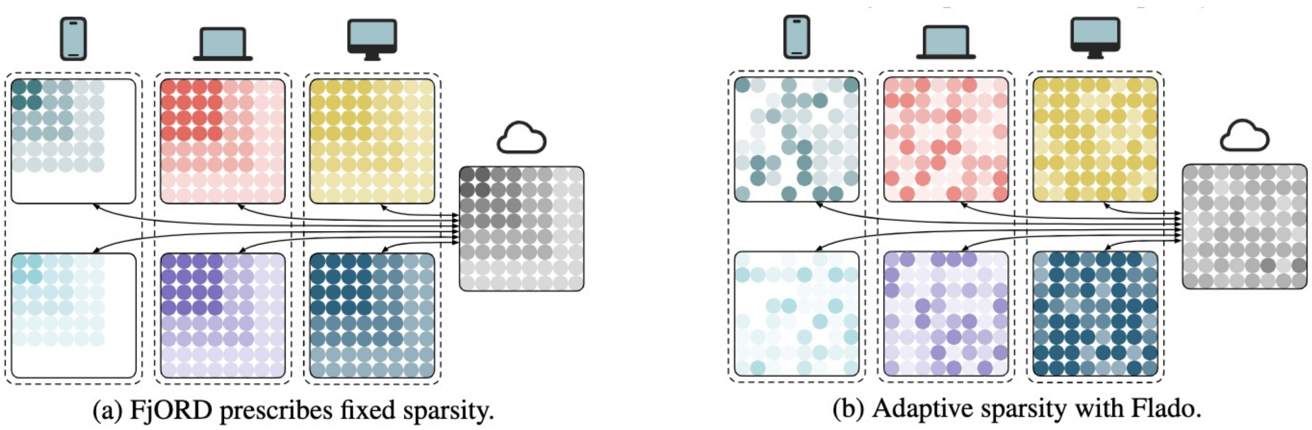


Figure 1. Comparing FjORD and the proposed method *Flado.*

Background of Federated Learning

Figure 2. European GDPR legislation

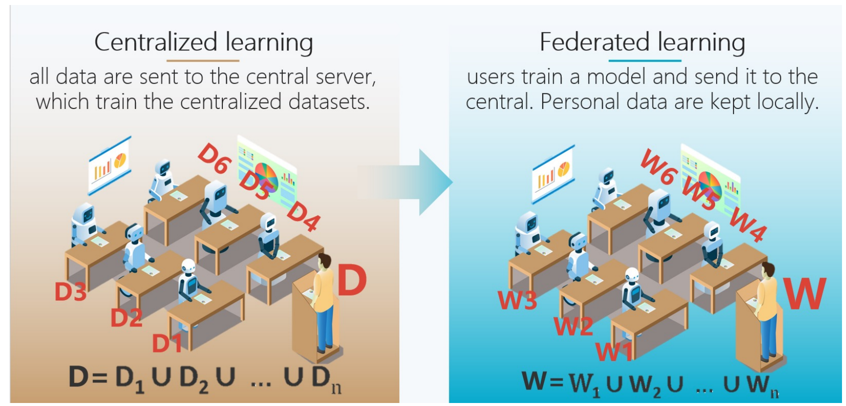


Figure 3. Comparing centralized and federated learning.

Due to incresingly stringent privacy protection legislations, the traditional centralized data analysis is no longer applicable for data located on massive edge devices.

image source: https://www.mn.uio.no/ifi/studier/masteroppgaver/nd/new-aggregation-methods-in-federated-learning.html

Addressing System Heterogeneity in FL

System Heterogeneity

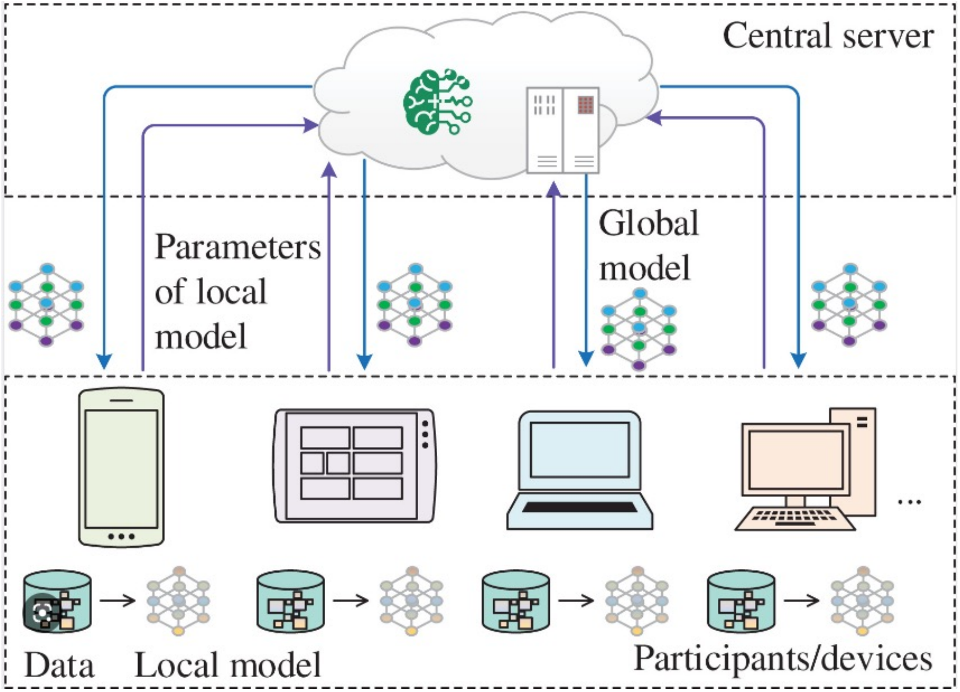


Figure 4. Participating clients may have different computing capabilities.

Addressing System Heterogeneity in FL

Existing works

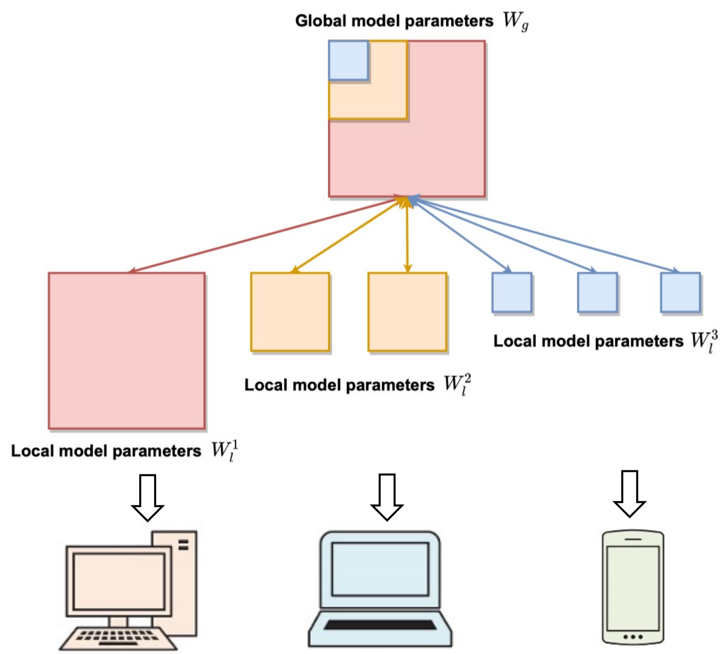


Figure 5. A concept illustration of HeteroFL[1]

HeteroFL: Slices submodels to adpat to devices with different computing capabilities

[1] Diao, Enmao et al. "HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients", ICLR2021

Addressing System Heterogeneity in FL

Existing works

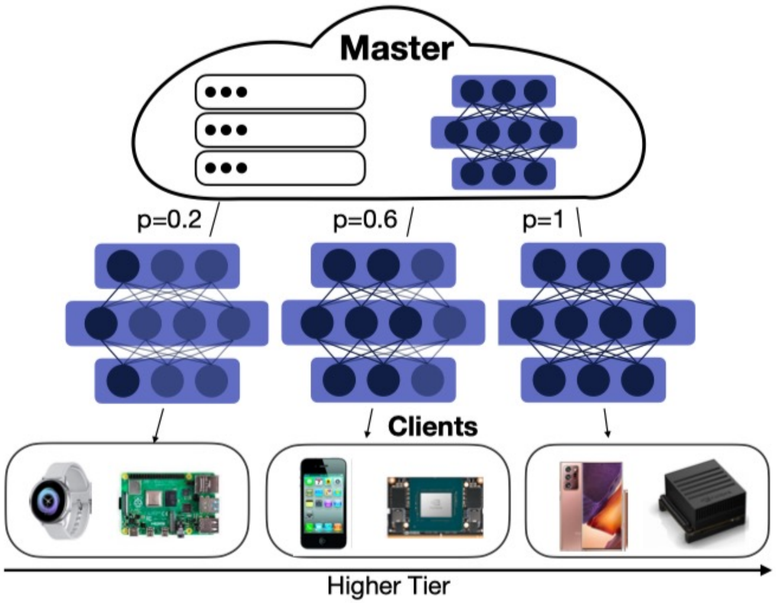


Figure 6. A concept illustration of FjORD[1]

FjORD: customizes the maximal model width and applies ordered dropout in each training step.

[1] Horvath et al. "Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout.". NeurIPS2021.

Addressing System Heterogeneity in FL

Limitations of existing works

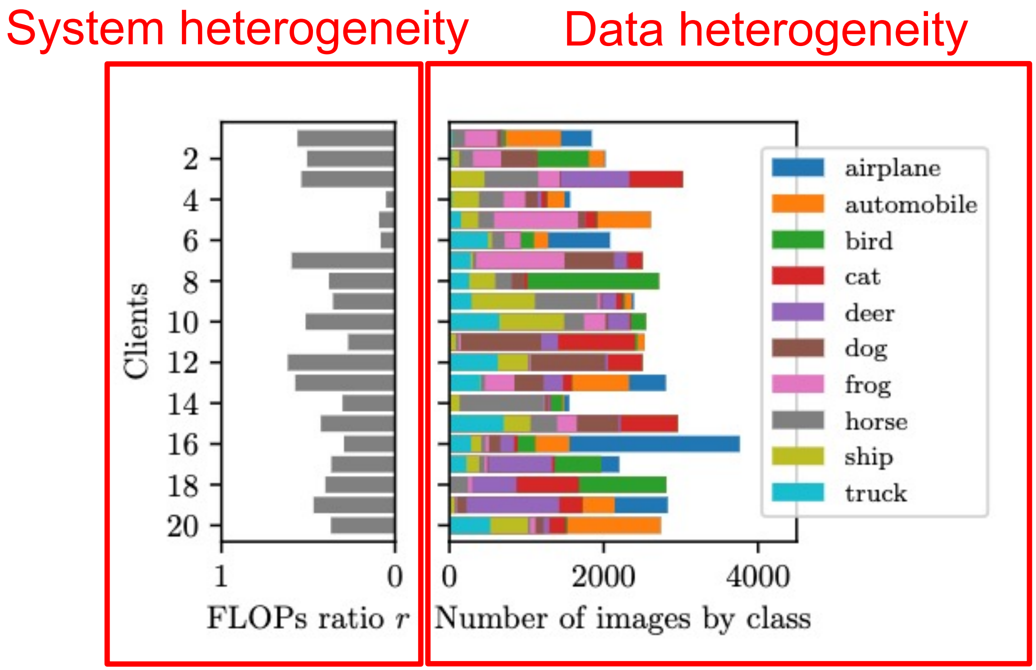


Figure 7. Existing works focus on system heterogeneity, but ignore the impact of local data distribution.

Limitations:

- Existing works prescribe a *coordinate-wise* sparsity pattern but ignored different data distributions among clients, which may cause conflicting gradient updates
- A fixed sparsity scheme could hinder collaborative training among clients, as some neurons are deactivated permanently.

Addressing System Heterogeneity in FL

Intuition: neurons may specilize to distinct features

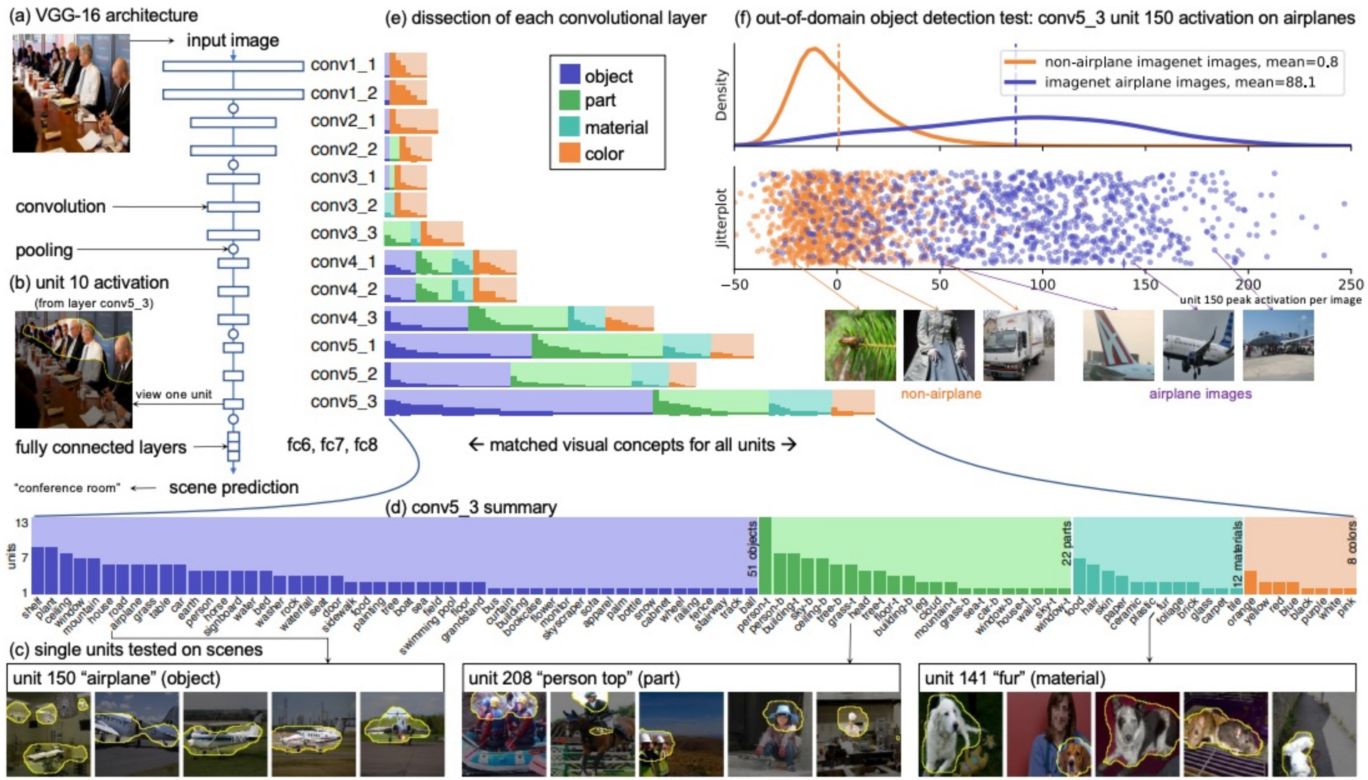


Figure 8. An illustration of single neuron activation on different objects within a VGG-16 scene classifier[1][2]

[1] Bau David et al. "Understanding the role of individual units in a deep neural network." *Proceedings of the National Academy of Sciences*, 2020.
[2] Zhou Bolei et al. Interpreting deep visual representations via network dissection. IEEE transactions on pattern analysis and machine intelligence, 2018.

Addressing System Heterogeneity in FL

Observation: local gradient update is contingent on data distribution

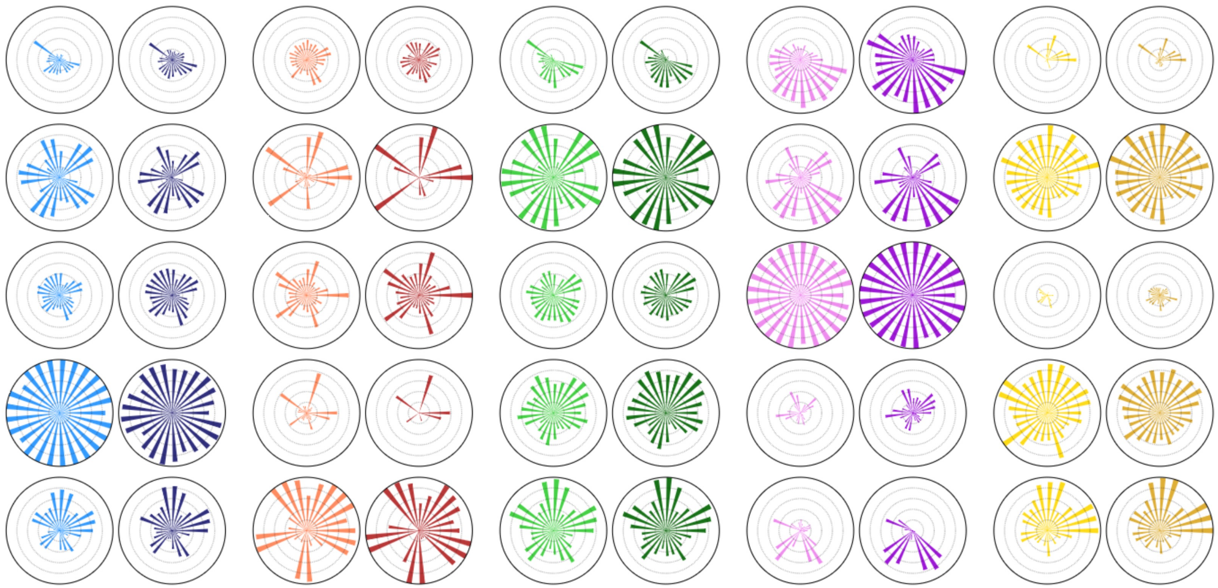


Figure 9. Similarity matrix of clients' gradient update direction

Observation: The clients allocated with the same digits shared similar update patterns, while different client pairs update direction is quite dispersed.

Addressing System Heterogeneity in FL

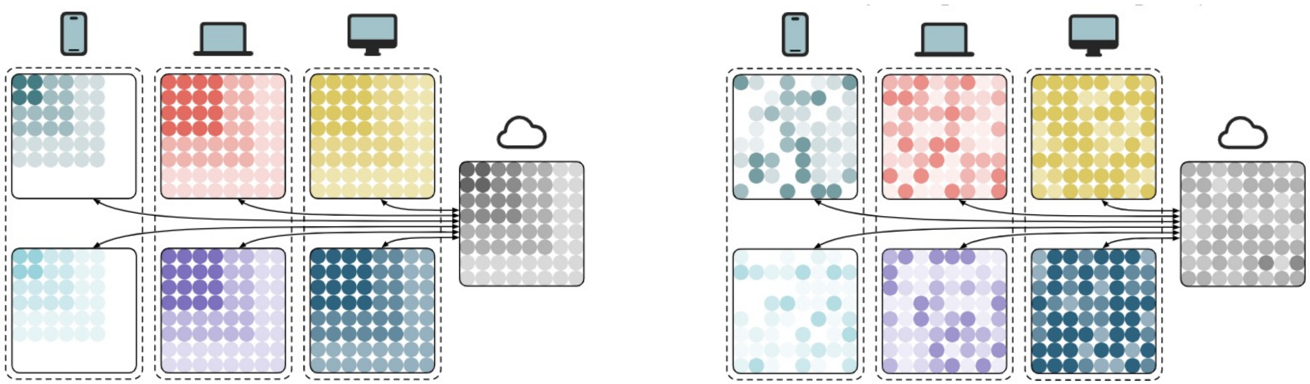
Motivation: can we foster collaboration by tailoring sparsity for each client?

Insights: It turns out that clients that shared similar data distributions tend to have similar updates, and by contrast different data distributions resulted in disparate updates.

Motivation: Can we concentrate training effort on neurons that specialize to the data distribution of the client, while paying less attention to neurons that are less relevant to the client?

Adaptive Channel Sparsity for Federated Learning

Proposed Method *Flado*: Federated Learning with Adaptive Dropout



(a) FjORD prescribes fixed sparsity.

(b) Adaptive sparsity with Flado.

Figure 10. Comparing FjORD and the proposed method *Flado.*

How to design the adaptive channel sparsity?

Challenges:

- pruning channel neurons would cause them to make no contribution later.
- prescribing a fixed sparsity scheme to channels can be suboptimal
- data heterogeneity causes clients to specialize to train different neurons

Adaptive Channel Sparsity for Federated Learning

Sparsity-driven Trajectory Alignment

$$J(\mathbf{z}) = \mathbf{PHDz},$$

fast Johnson-Lindenstrauss transform (FJLT)

$$\max_{\mathbf{p}_c} \mathbb{E}_{\mathbf{b}_c \sim \mathcal{B}(\mathbf{p}_c)}$$

$$\text{cossim}\left(J\left(\Delta\boldsymbol{\theta}^{(t)}\right), J\left(\nabla_{\boldsymbol{\theta}^{(t)}} \ell_c\left(\mathbf{b}_c \circ \boldsymbol{\theta}^{(t)}\right)\right)\right),$$

$$s.t.\ g_c(r_c, \mathbf{p}_c) \geq 0.$$ FLOPs budget constraint

FLOPs constraint

$$g_c(r_c, \mathbf{p}_c) = r_c - \mathrm{flops}(\ell_c, \mathbf{p}_c) / \mathrm{flops}(\ell_c, \mathbf{1}),$$

Adaptive Channel Sparsity for Federated Learning

Main Results

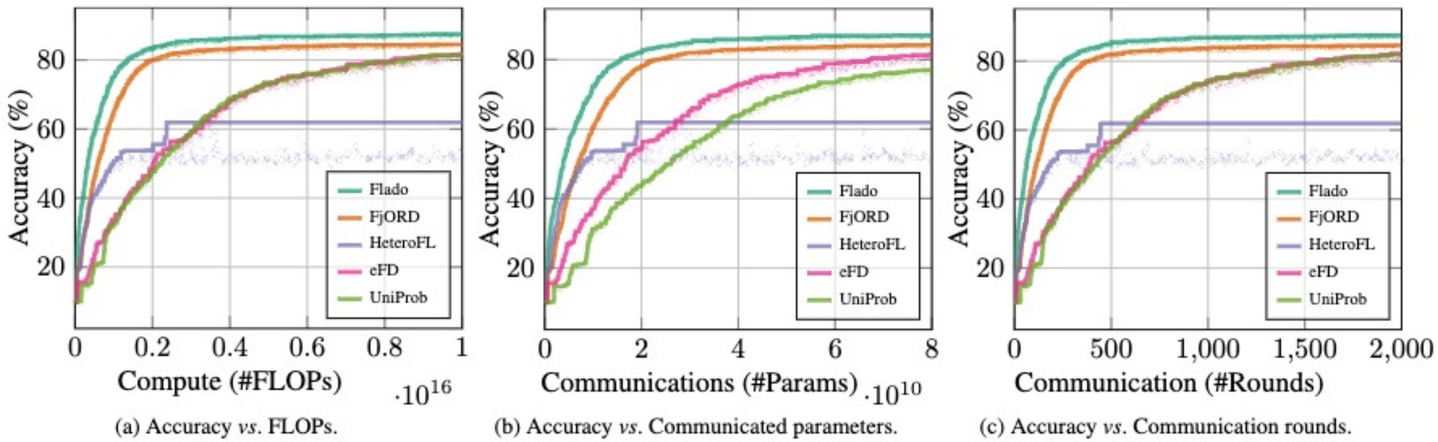


Figure 11. Comparison of convergence curves on CIFAR10.

Flado attains consistently higher converged accuracies.

Adaptive Channel Sparsity for Federated Learning

Main Results

Table 1. Comparing the sparse FL algorithms on converged accuracies, computation and communication costs.

Method	CIFAR-10			Permitting −5% accuracy budget from Flado			Permitting −10% accuracy budget from Flado		
	Accuracy	Δ FLOPs	Δ Comm. Params	Rounds	FLOPs	CommParams	Rounds	FLOPs	Comm. Params
HeteroFL	53.83%±3.66%	2.13 P±14.67 T	18.08 G±124.45 M	—	—	—	—	—	—
UniProb	80.91%±0.61%	−2.90 P±2.12 P	−14.01 G±18.03 G	—	—	—	1329.5±2.9	6.21 P±14.48 T	75.59 G±176.28 M
eFD	81.82%±0.31%	−113.08 T±91.82 T	−34.02 G±1.09 G	—	—	—	1287.5±5.2	6.01 P±25.31 T	51.56 G±218.57 M
FjORD	84.38%±0.18%	−6.59 P±76.82 T	−47.06 G±657.26 M	562.5±5.76	2.63 P±14.48 T	31.98 G±176.28 M	314.5±4.6	1.47 P±22.77 T	17.88 G±277.20 M
Flado	**87.24%**±0.17%	−7.21 P±6.18 T	−87.71 G±75.17 M	**330.5**±3.45	**1.54 P**±8.99 T	**18.79 G**±108.93 M	**215.5**±2.3	**1.01 P**±11.71 T	**12.25 G**±142.62 M

Method	SVHN			Permitting −2% accuracy budget from Flado			Permitting −5% accuracy budget from Flado		
	Accuracy	Δ FLOPs	Δ Comm. Params	Rounds	FLOPs	Comm. Params	Rounds	FLOPs	Comm. Params
HeteroFL	89.07%±0.23%	390.35 T±1.02 T	38.63 G±103.75 M	—	—	—	163.5±1.7	55.28 T±647.74 G	5.47 G±64.11 M
UniProb	90.39%±0.07%	−299.57 T±83.02 T	−22.48 G±8.22 G	—	—	—	426.5±2.9	139.25 T±1012.26 G	18.07 G±131.38 M
eFD	91.11%±0.06%	−226.29 T±40.39 T	−39.24 G±5.24 G	1540.5±2.3	502.85 T±804.65 G	51.35 G±83.41 M	430.5±5.2	140.55 T±1.75 T	14.35 G±182.26 M
FjORD	92.36%±0.04%	−399.73 T±10.61 T	−34.31 G±1.08 G	667.5±3.5	217.86 T±1.18 T	28.29 G±156.45 M	253.5±1.7	82.75 T±625.16 G	10.74 G±81.18 M
Flado	**92.90%**±0.04%	−354.03 T±1.46 T	−45.98 G±194.05 M	**442.5**±2.9	**144.48 T**±1012.26 G	**18.75 G**±131.38 M	**199.5**±2.9	**65.14 T**±1012.26 G	**8.45 G**±131.38 M

Method	Fashion-MNIST			Permitting −5% accuracy budget from Flado			Permitting −10% accuracy budget from Flado		
	Accuracy	Δ FLOPs	Δ Comm. Params	Rounds	FLOPs	Comm. Params	Rounds	FLOPs	Comm. Params
UniProb	83.00%±0.11%	1.83 P±2.84 T	44.38 G±69.06 M	—	—	—	698.5±1.7	656.28 T±1.76 T	15.56 G±42.67 M
eFD	84.94%±0.09%	−1.18 P±3.66 T	−33.68 G±88.83 M	—	—	—	410.5±2.3	386.11 T±2.29 T	6.24 G±38.14 M
FjORD	85.54%±0.06%	−253.89 T±51.17 T	+7.49 G±847.86 M	—	—	—	366.5±1.1	344.15 T±1.21 T	8.16 G±29.45 M
HeteroFL	87.29%±0.17%	−1.35 P±3.66 T	−36.75 G±88.83 M	498.5±4.6	491.32 T±4.69 T	7.66 G±74.94 M	122.5±5.8	115.10 T±5.56 T	2.73 G±134.95 M
Flado	**90.58%**±0.09%	−1.05 P±147.24 T	−11.56 G±2.30 G	**354.5**±4.0	**333.07 T**±3.93 T	**7.90 G**±95.42 M	**81.5**±1.7	**80.33 T**±1.84 T	**1.25 G**±29.45 M

Flado is more efficient than competing methods on computation and communication cost.

Adaptive Channel Sparsity for Federated Learning

Main Results

Table 2. Comparing the sparse FL algorithms under increasing level of data heterogeneity.

$\alpha = \infty$	Accuracy	Δ FLOPs	Δ Comm. Params
HeteroFL	83.20%±0.42%	2.49 P±14.67 T	21.13 G±124.45 M
UniProb	85.38%±0.22%	+0.80 P± 2.02 P	+31.49 G± 17.14 G
eFD	85.86%±0.21%	−0.87 P±98.72 T	−40.32 G± 1.17 G
FjORD	87.58%±0.09%	−6.43 P±36.67 T	−44.91 G±313.70 M
Flado	**89.16**%±0.08%	−7.13 P±58.68 T	−86.80 G±714.29 M

$\alpha = 5$	Accuracy	Δ FLOPs	Δ Comm. Params
HeteroFL	82.51%±0.34%	5.17 P±14.67 T	43.86 G±124.45 M
UniProb	84.82%±0.17%	+1.69 P±14.67 T	+39.67 G±124.45 M
eFD	85.69%±0.25%	−1.75 P±14.48 T	−48.29 G±176.28 M
FjORD	86.92%±0.17%	−5.38 P±14.47 T	−32.17 G±123.96 M
Flado	**88.85**%±0.10%	−6.97 P±14.48 T	−84.81 G±176.28 M

$\alpha = 0.05$	Accuracy	Δ FLOPs	Δ Comm. Params
HeteroFL	28.06%±5.04%	2.27 P±14.67 T	19.29 G±124.45 M
UniProb	63.05%±1.14%	+0.60 P±14.67 T	+15.51 G±124.45 M
eFD	62.84%±1.20%	−0.34 P±13.10 T	−49.88 G±159.45 M
FjORD	77.64%±0.91%	−10.54 P±14.44 T	−82.39 G±124.11 M
Flado	**79.14**%±1.12%	−5.92 P±14.48 T	−72.07 G±176.28 M

Flado tolerates aggressive data heterogeneity.

Adaptive Channel Sparsity for Federated Learning

Main Results

Table 3. Comparing the sparse FL algorithms under increasing level of system heterogeneity.

$\mathcal{U}(0.64, 0.64)$	Accuracy	Δ FLOPs	Δ Comm. Params
FjORD	87.01% ±0.11%	15.26 P ±23.83 T	112.89 G ±176.28 M
HeteroFL	87.43% ±0.09%	−4.72 P ±24.22 T	−47.32 G ±150.50 M
UniProb	87.44% ±0.13%	−3.11 P ±23.79 T	−4.81 G ±176.28 M
eFD	88.26% ±0.09%	−2.18 P ±23.82 T	−29.42 G ±149.96 M
Flado	**88.82%** ±0.14%	−11.88 P ±23.83 T	−25.67 G ±176.28 M

$\mathcal{U}(0.32, 0.64)$	Accuracy	Δ FLOPs	Δ Comm. Params
HeteroFL	58.91% ±4.23%	2.96 P ±19.11 T	21.40 G ±138.38 M
eFD	85.55% ±0.15%	+227.50 T ±18.92 T	+1.89 G ±138.11 M
UniProb	86.08% ±0.31%	−2.81 P ±18.84 T	−1.70 G ±176.28 M
FjORD	86.43% ±0.14%	−102.70 T ±18.84 T	−960.70 M ±176.28 M
Flado	**87.88%** ±0.13%	−8.98 P ±18.84 T	+3.04 G ±176.28 M

$\mathcal{U}(0.16, 0.64)$	Accuracy	Δ FLOPs	Δ Comm. Params
HeteroFL	57.64% ±4.65%	2.52 P ±16.53 T	20.02 G ±131.09 M
eFD	83.82% ±0.28%	+842.98 T ±16.44 T	+6.78 G ±131.05 M
UniProb	83.89% ±0.43%	−331.42 T ±16.35 T	+25.80 G ±176.28 M
FjORD	85.84% ±0.19%	−6.83 P ±16.35 T	−73.60 G ±176.28 M
Flado	**86.91%** ±0.16%	−5.09 P ±16.34 T	−54.90 G ±176.28 M

$\mathcal{U}(0.08, 0.64)$	Accuracy	Δ FLOPs	Δ Comm. Params
HeteroFL	57.39% ±4.20%	2.35 P ±15.36 T	19.39 G ±126.96 M
eFD	81.45% ±0.59%	+999.93 T ±15.02 T	+8.46 G ±126.61 M
UniProb	81.70% ±0.52%	−186.77 T ±15.10 T	+29.60 G ±176.28 M
FjORD	84.58% ±0.19%	−6.62 P ±15.10 T	−77.24 G ±176.28 M
Flado	**86.98%** ±0.11%	−5.93 P ±15.08 T	−69.17 G ±176.28 M

Flado is highly elastic under different system heterogeneity levels.

Thanks