UC San Diego

THE WEB CONFERENCE ACM

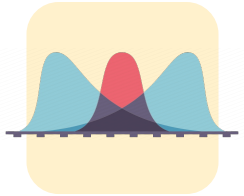
How Few Davids Improve One Goliath: Federated Learning in Resource-Skewed Edge Computing Environments

Jiayun Zhang, Shuheng Li, Haiyu Huang, Zihan Wang, Xiaohan Fu, Dezhi Hong,
Rajesh K. Gupta, Jingbo Shang
University of California, San Diego

Background: Heterogeneous Computing Environments

- People and organizations interact with digital systems in distributed environments, which demonstrate diversity and variation.
- Sources of Heterogeneity:

Devices:
e.g. computers, iPads, mobile phones, wearable sensors…



Data:
user-specific data e.g. content preference

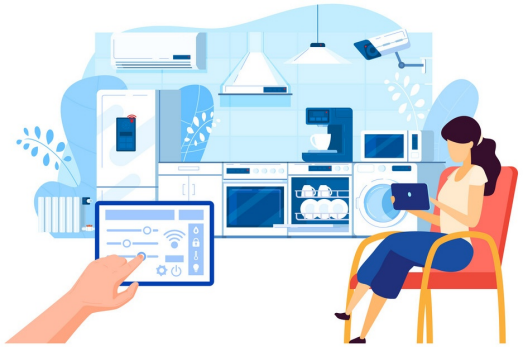
Heterogeneity emphasizes privacy concerns!

Background: Federated Learning

- **Federated Learning**
  - Multiple actors collaboratively train models in distributed environments without sharing data.
- **Applications**
  - healthcare, mobile computing, smart home.

Healthcare
(e.g., medical diagnosis)

Mobile computing
(e.g., human activity recognition)



Smart home
(e.g., voice assistants)

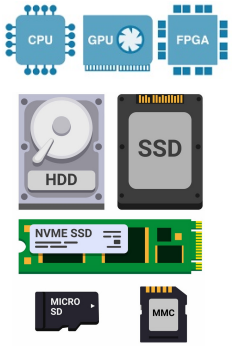
System Heterogeneity: Resource Skew

Diverse Edge Devices

- Enterprise-owned computing machines
- User-owned mobile/WoT devices

Different Device Configurations

- Processor
- Memory
- Data storage (I/O)
- …

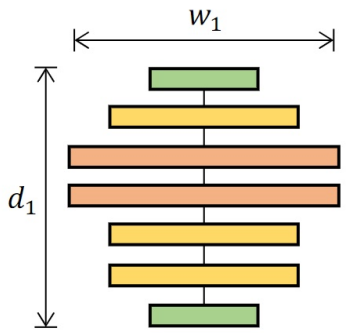


Resource Skew

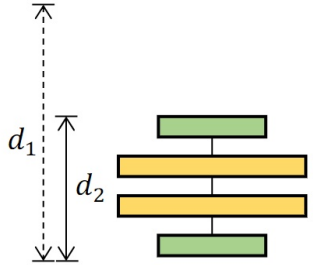
- Few (1-2) strong devices and many weak devices.
- e.g., a single company trains ML models with its end users using federated learning.

Existing methods

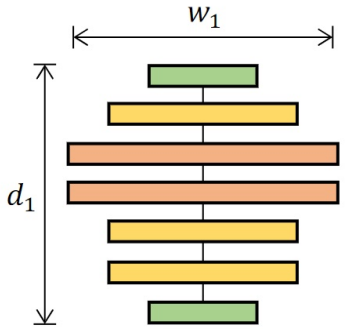
- Federated learning with heterogeneous models.
- Downscale the large model to fit varying device capacities.



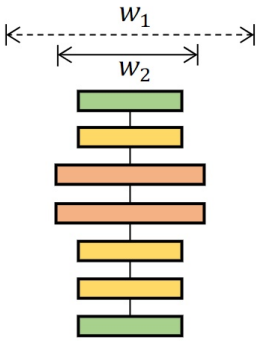
(a) original model



(b) depth-wise



(a) original model

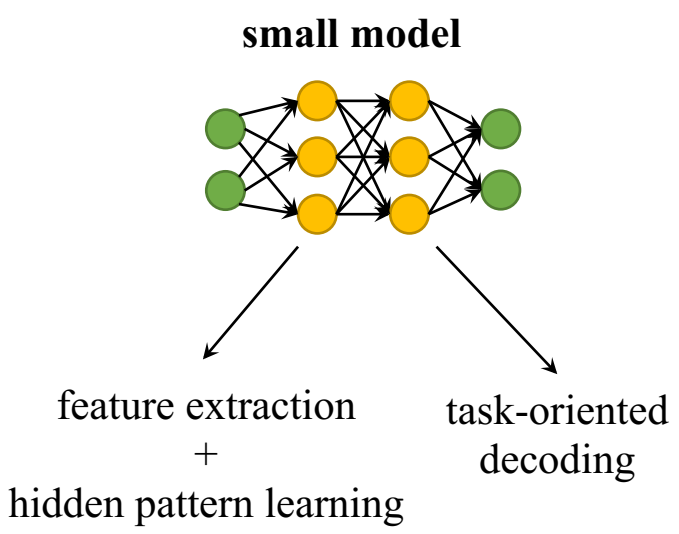


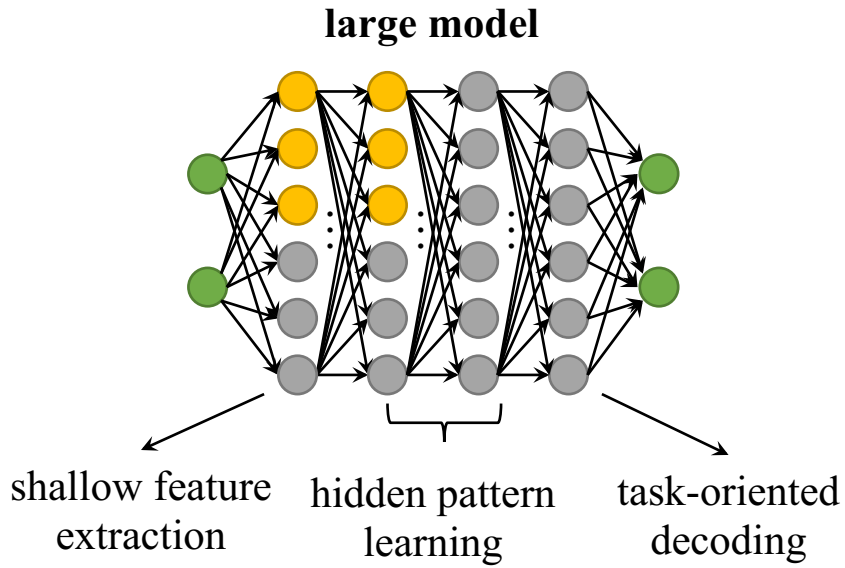
(c) width-wise

- Identify and aggregate the common parts by weight averaging.

Challenges

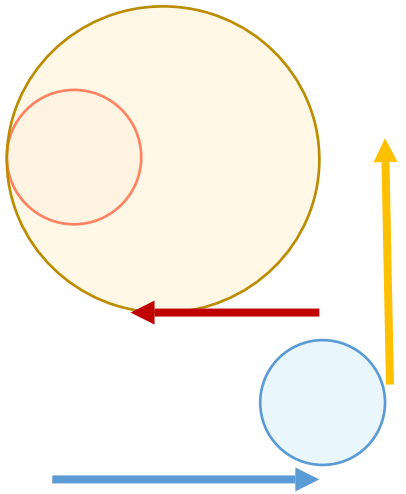
- The unshared portion (grey) of the large model rarely receives updates or derives benefits from others.
- The functionalities of layers/neurons are misaligned.





Performance of Prior Work with Resource Skew

- Small models are insufficient for strong devices
  - AllSmall v.s. ExclusiveFL
- Weak devices benefit from collaboration with strong devices
  - ExclusiveFL v.s. other methods on weak devices
- Strong devices derive minimal benefits from weak devices with existing methods
  - ExclusiveFL v.s. other methods on strong devices



AllSmall: training small models on all devices.
ExclusiveFL: strong devices run large models and weak devices run small models and conduct federated learning within their respective groups.

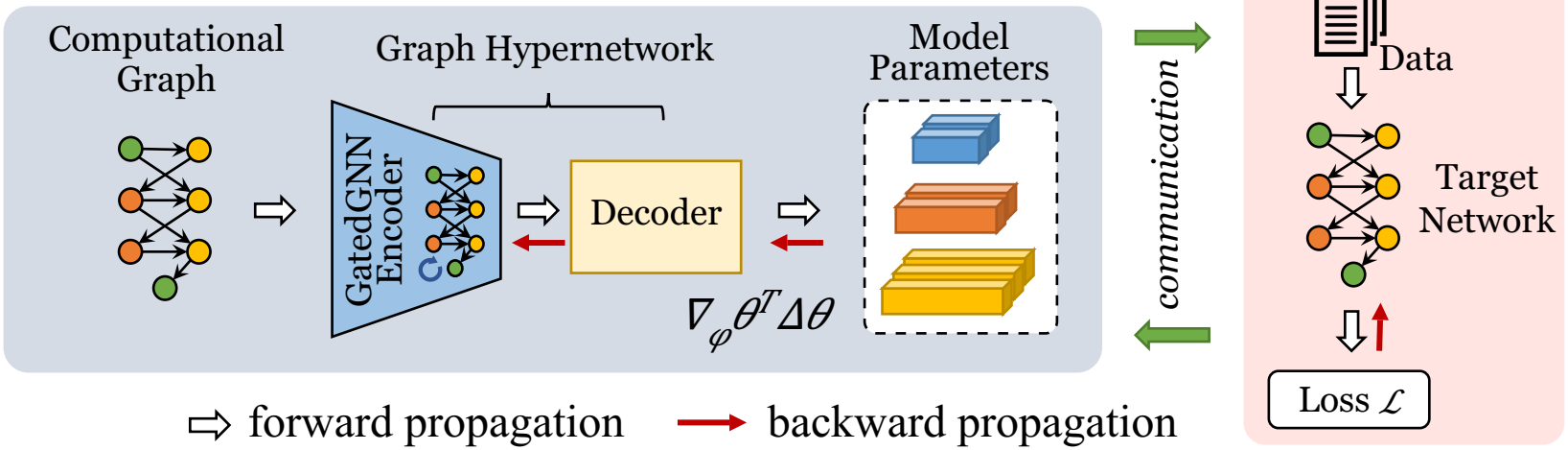
Method: RecipFL Overview

- **Research Question:** *Can strong devices benefit from weak devices in resource-skewed environments?*
- Train a graph hypernetwork to generate weights for personalized client models.
- Agnostic to model scaling strategies and enable collaboration among arbitrary neural network models.

Method: Graph Hypernetwork

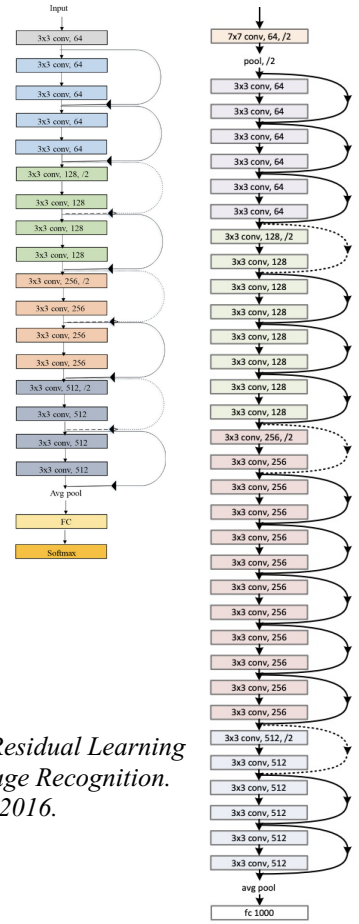
- A neural network that generates model weights for another network.

$$\phi \leftarrow \phi - \eta_s \sum_{m \in S_t} (\nabla_\phi \theta_m)^T \Delta\theta_m$$



Training Graph Hypernetwork in *Federated* Setting

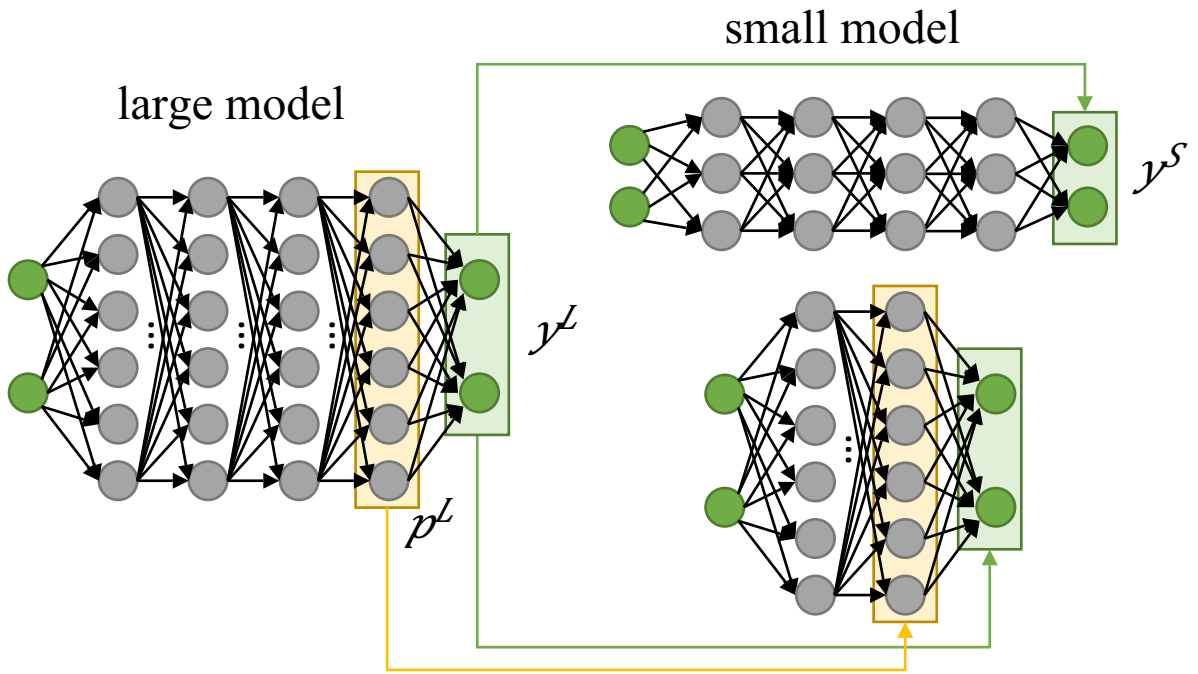
ResNet-18 ResNet-34



Deep Residual Learning for Image Recognition. CVPR 2016.

Method: Knowledge Transfer from Strong to Weak Devices

- Generate and train both small and large models on strong devices
- Knowledge distillation from large to small model
  - Cross entropy loss: let the small model mimic the prediction probabilities of the large model
  - KL-divergence loss: align the feature spaces between the small and large models



$CE(y^S, y^L)$

$D_{KL}(p^S, p^L)$

Experiment

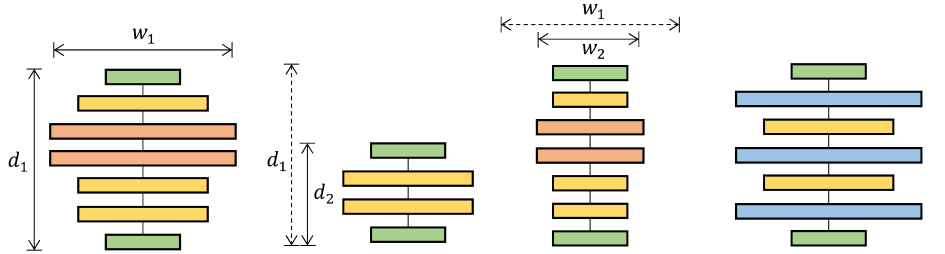
- Image and language classification tasks
- Multiple network architectures

Table 2: Federated learning configurations.

Dataset	# of devices		Data allocation		Large model	# of parameters			Pretrained?
	Strong	Weak	Strong	Weak		Original	Depth-scaled	Width-scaled	
CIFAR-10	1	5	50%	10%	ResNet-18	11M	450K	444K	✗
CIFAR-100	1	50	50%	1%	DenseNet-121	1M	258K	276K	✗
MNIST	2	500	25%	0.1%	LeNet-5	44K	5.6K (scaled in depth & width)		✗
MNLI	1	20	50%	2.5%	BERT	110M	67M (DistilBERT)		✓

Table 1: RecipFL is compatible with various ways of model scaling, showing more flexibility than existing solutions.

Scaling Strategy	HeteroFL	InclusiveFL	FlexiFed	DepthFL	RecipFL (ours)
depth-wise		✓	✓	✓	✓
width-wise	✓				✓
architecture-wise					✓



(a) original model (b) depth-wise (c) width-wise (d) architecture-wise

Experiment: Main results

- RecipFL consistently outperforms the compared methods, regardless of model scaling strategies and NN architectures.
- RecipFL improves the model performance on both strong and weak devices.

Table 3: Experiment results (average accuracy and standard deviation). RecipFL consistently outperforms the compared methods across all datasets and model scaling strategies, benefiting both strong and weak devices.

Scaling	Method	CIFAR-10		CIFAR-100	
		Strong	Weak	Strong	Weak
Depth	AllSmall	64.34±2.14	68.50±3.42	17.86±2.56	25.56±3.11
	ExclusiveFL	84.85±1.85	59.11±4.22	32.21±3.81	19.22±2.07
	FlexiFed [38]	82.86±1.77	67.66±3.93	28.60±3.48	27.84±2.98
	InclusiveFL [29]	83.22±0.47	67.66±3.14	18.98±3.49	28.71±2.87
	DepthFL [16]	73.90±1.49	78.16±1.48	22.08±3.58	36.83±2.87
	RecipFL	**85.28±0.22**	**78.65±1.35**	**41.63±2.24**	**45.52±3.12**
Width	AllSmall	82.86±1.77	78.90±2.87	29.80±3.32	37.90±2.83
	ExclusiveFL	83.96±1.97	70.65±3.99	32.22±6.66	24.49±3.52
	HeteroFL [6]	84.76±1.19	77.93±2.92	26.51±2.70	39.05±2.82
	RecipFL	**85.06±0.13**	**82.88±1.29**	**43.64±2.84**	**42.00±3.88**

Method	MNIST		MNLI	
	Strong	Weak	Strong	Weak
AllSmall	91.73±3.94	77.05±7.47	73.47±0.52	82.13±2.89
ExclusiveFL	92.97±1.98	77.85±5.14	80.20±0.20	70.52±6.04
FlexiFed [38]	92.70±2.72	73.89±8.08	79.65±0.18	82.31±6.15
InclusiveFL [29]	84.77±3.12	75.72±6.27	79.87±0.30	81.17±4.31
DepthFL [16]	94.33±1.95	78.39±7.52	77.11±0.90	80.92±6.64
HeteroFL [6]	89.53±3.22	75.95±8.01	79.65±0.18	82.31±6.15
RecipFL	**97.07±1.87**	**86.36±6.60**	**82.78±0.57**	**83.37±4.72**

Experiment: Miscellany

- RecipFL is not limited to the setup of one large and one small model architecture and can work with diverse device capacities.

Table 4: Performance with more diverse device capacities.

Method	Small (LeNet-5)	Medium (ResNet-101)	Large (VGG-16)
AllSmall	69.22±2.16	66.58±1.84	59.06±1.41
ExclusveFL	46.00±3.71	72.38±3.80	79.86±2.52
RecipFL	**70.12±2.33**	**86.37±1.72**	**81.23±0.41**

- We also conduct ablation study and experiment on the impact of different resource skewness settings, e.g., data ratio and number of strong or weak devices.

Contribution Summary

- Study federated learning with resource skew
- Propose a new framework: RecipFL
  - Graph hypernetwork at the server - producing personalized weights for clients
- Achieve state-of-the-art performance
  - Weak devices can also contribute meaningfully to the learning system

Thank you!

Q&A

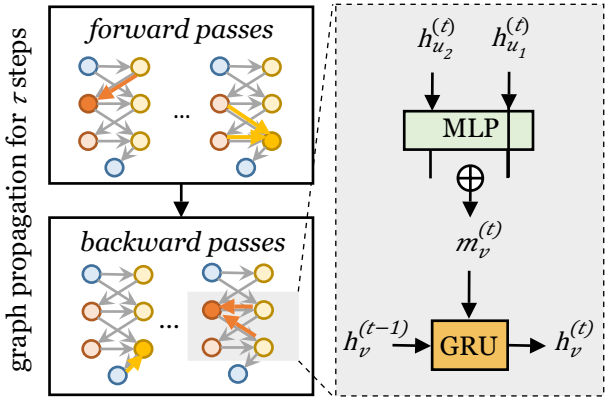
Method: Weight Generation with Graph Hypernetwork

Encoding:

- Use GatedGNN to encode node features.
- Topologically traverses the nodes.
- Message passing and node feature update.

$$m_v^{(t)} = \sum_{u \in \mathrm{IN}(v)} \mathrm{MLP}(h_u^{(t)})$$

$$h_v^{(t)} = \mathrm{GRU}(h_v^{(t-1)}, m_v^{(t)})$$

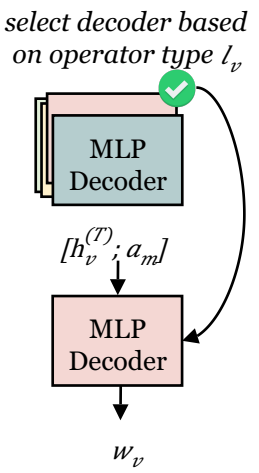


Method: Weight Generation with Graph Hypernetwork

Decoding:

- Each type of parametric layer (e.g., conv, linear) uses an MLP to decode weights.
- Client descriptors: describe the class distribution of every client.
- Combine node embedding and client descriptor as input.

$$w = \{w_v | v \in \mathcal{V}\} = \{\mathrm{MLP}_{l_v}([h_v^{(\tau)}, a_m]) | v \in \mathcal{V}\}$$



Experiment: Ablation Study

- Ablated version: RecipFL w/o knowledge transfer (from strong to weak)
- RecipFL w/o KT significantly improves over the naive baseline AllSmall
- RecipFL w/o KT can often outperform the state-of-the-art method DepthFL.

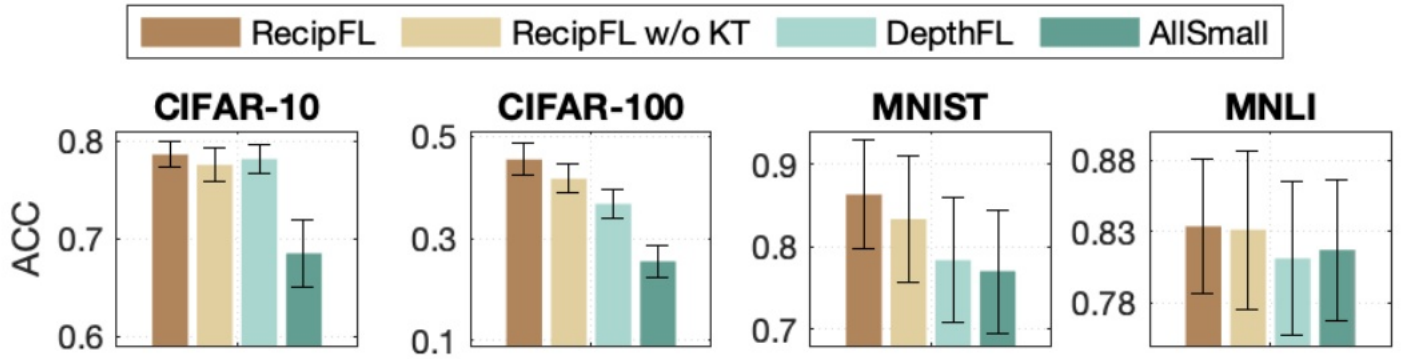
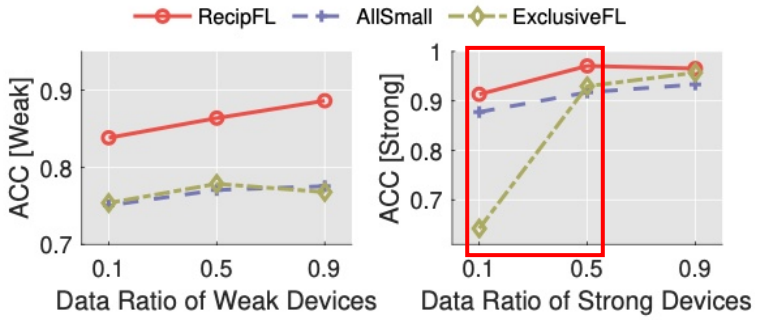


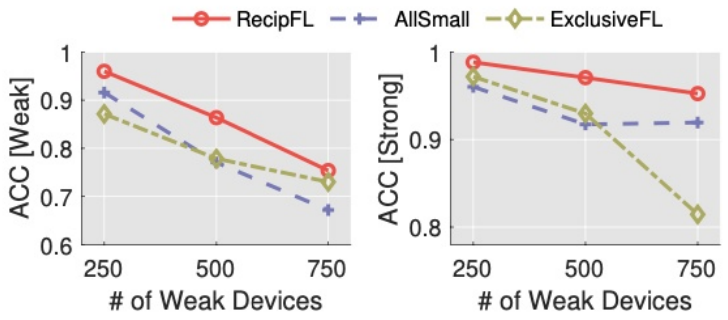
Figure 6: Ablation study: performance of weak devices.

Experiment: Exploratory Studies

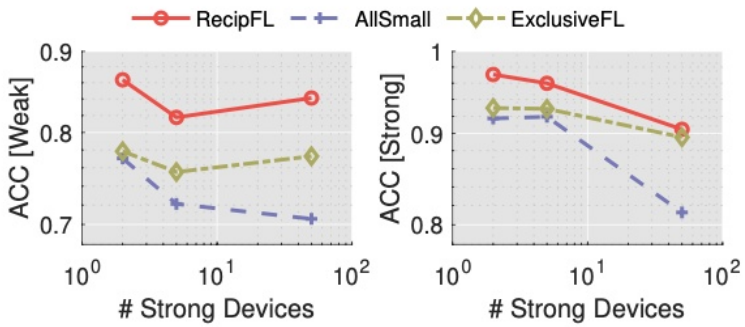
- RecipFL exhibits superior scalability and robustness across a range of resource skew scenarios, consistently enhancing the performance of both strong and weak devices.



(a) Impact of data ratio



(b) Varying number of weak devices



(c) Varying number of strong devices