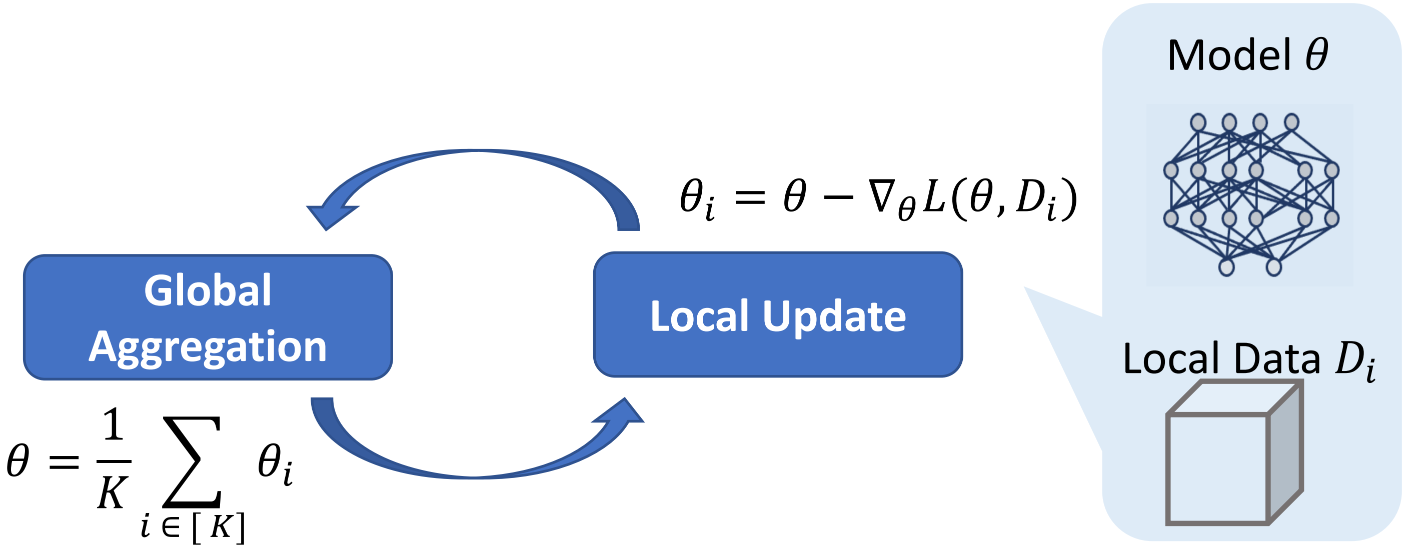
FedResCue: *Res*ilient and *C*ommu*n*ication-*E*fficient Learning for Heterogeneous *Fed*erated Systems

Zhuangdi Zhu, Junyuan Hong, Steve Drew, and Jiayu Zhou

Proceedings of the 39th International Conference on Machine Learning, 2022

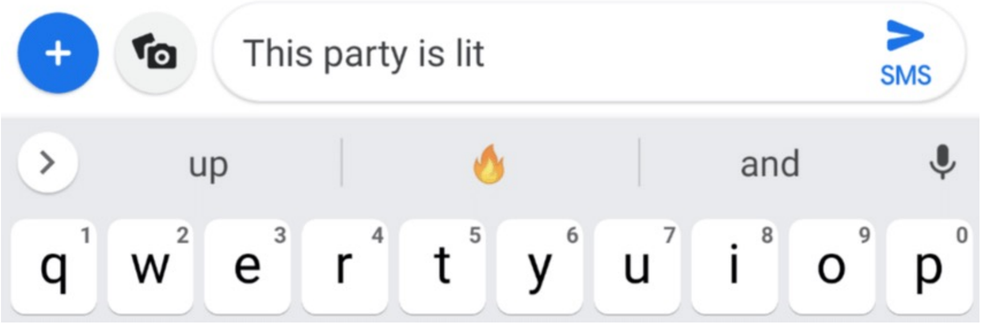
Federated Learning

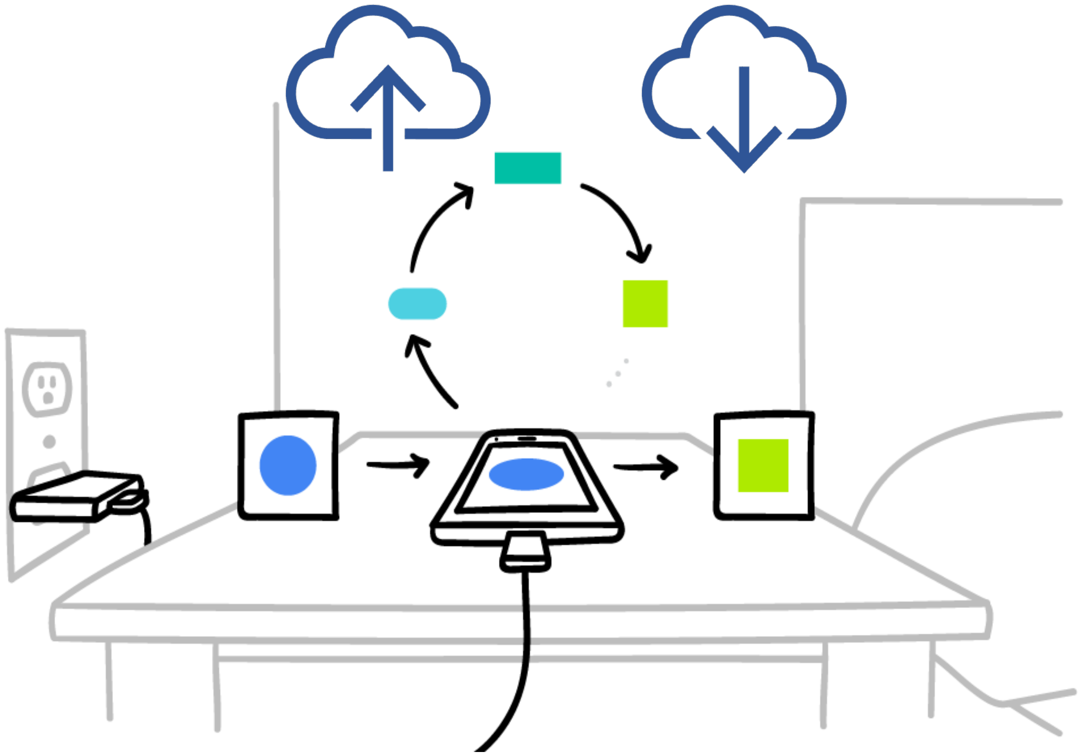
- A *decentralized* machine learning paradigm
  - Client: Local learning
  - Server: Global aggregation



Example of Federated Learning Application

- Natural Language Processing
  - *Gboard* service by Google



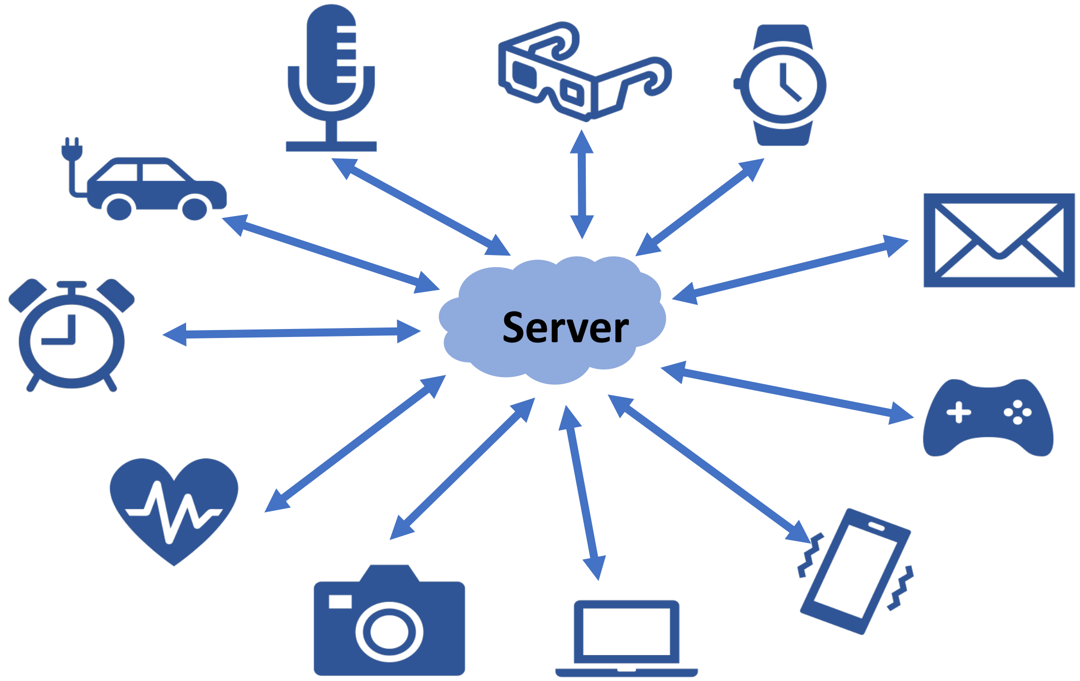


Source: Google AI blog

Challenges of Federated Learning

System Heterogeneity

- Clients diverge in ***memory*** and ***bandwidths*** capacities.



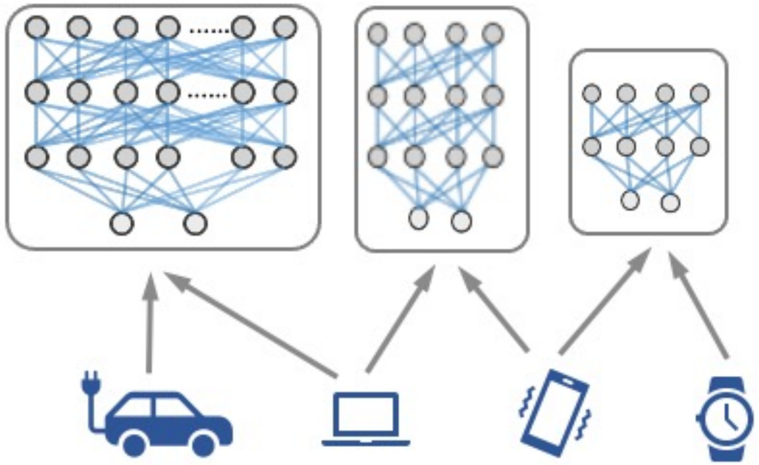
Challenges of Federated Learning

System Heterogeneity

- Traditional FL algorithms require unified model size for global aggregation:

$$\theta = \frac{1}{K} \sum_{i \in [K]} \theta_i$$

- One model architecture may not fit all clients.



Challenges of Federated Learning

Connection Uncertainty

- Network connections are noisy and ***unstable*** in real world.

  - ***Unreliable*** to transmit *large model* parameters

Challenges of Federated Learning

Connection Uncertainty

- Network connections are noisy and ***unstable*** in real world.

  - ***Unreliable*** to transmit *large model* parameters
- Dropped clients affect the global model quality:

$$\theta = \frac{1}{K} \sum_{i \in [K]} \theta_i$$

Paper Outline

Background and Challenges in Federated Learning

- System Heterogeneity
- Unstable connection

Motivation and Key idea

- Learning structurally prunable networks

Methodology

- Self-distilled network via progressive learning

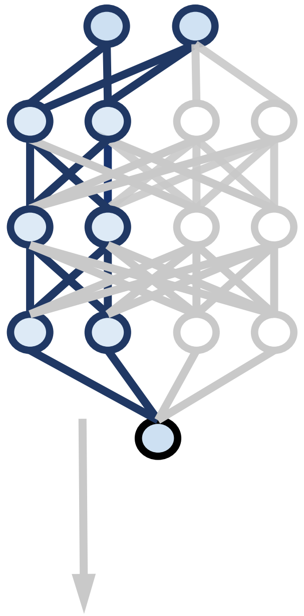
Performance Evaluation

- Robustness
- Communication Efficiency

Key Idea: *Learning* and Transmitting Structurally Prunable Models

During FL Local Learning:

- A model can by ***structurally pruned*** by removing its tailing channels at each layer
- A pruned sub-model shall be functional without the need of fine-tuning.

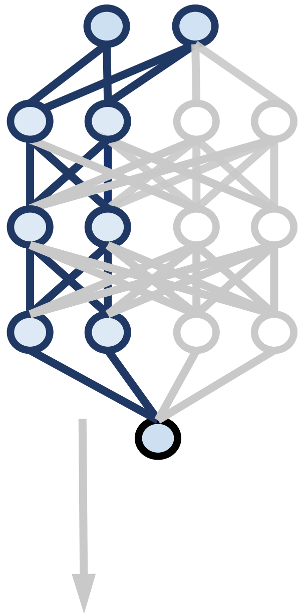


A structurally pruned sub-model

Key Idea: *Learning* and Transmitting Structurally Prunable Models

During FL Local Learning:

- Without loss of generality, we use a unified pruning ratio for all layers to prune a model.
- A sub-model is specified with a pruning ratio p
  - Which can be treated as a sequence of ***columns.***

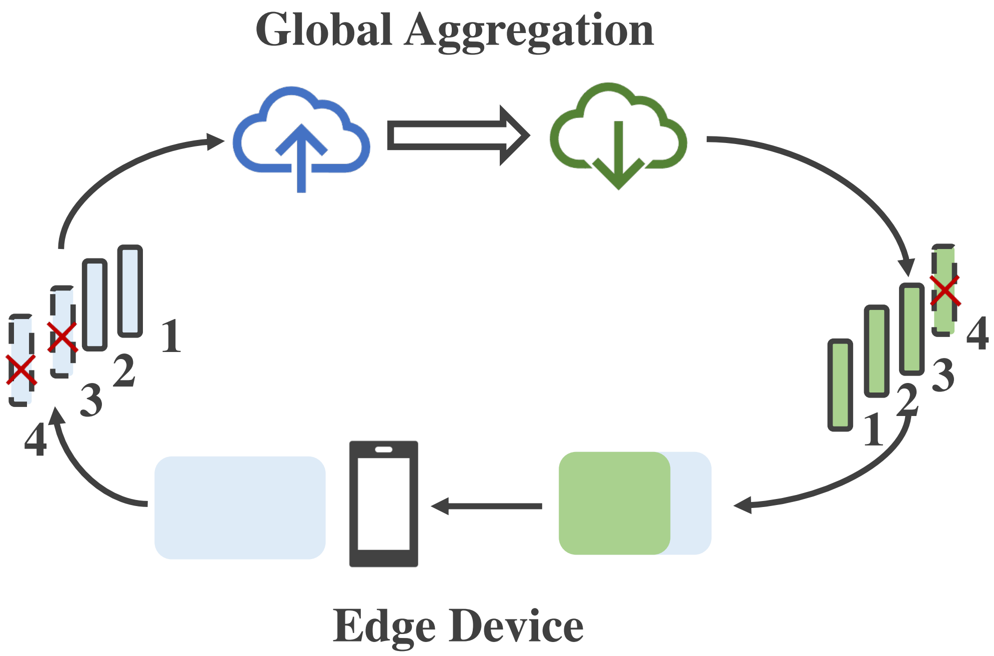


A pruned sub-model with pruning ratio $p = 0.5$

Key Idea: Learning and <u>*Transmitting*</u> Structurally Prunable Models

During FL Communication:

- Model parameters of columns are transmitted ***sequentially*** between the server and the client.



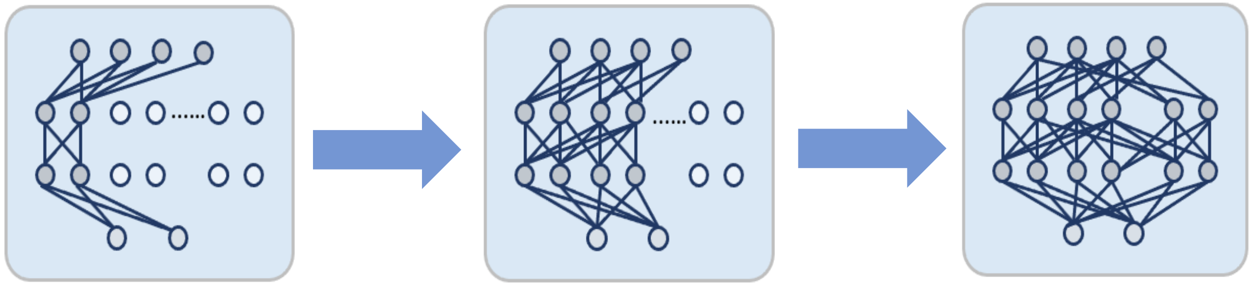
Sequential model transmission.

Key Idea: Learning and *Transmitting* Structurally Prunable Models

During FL Communication:

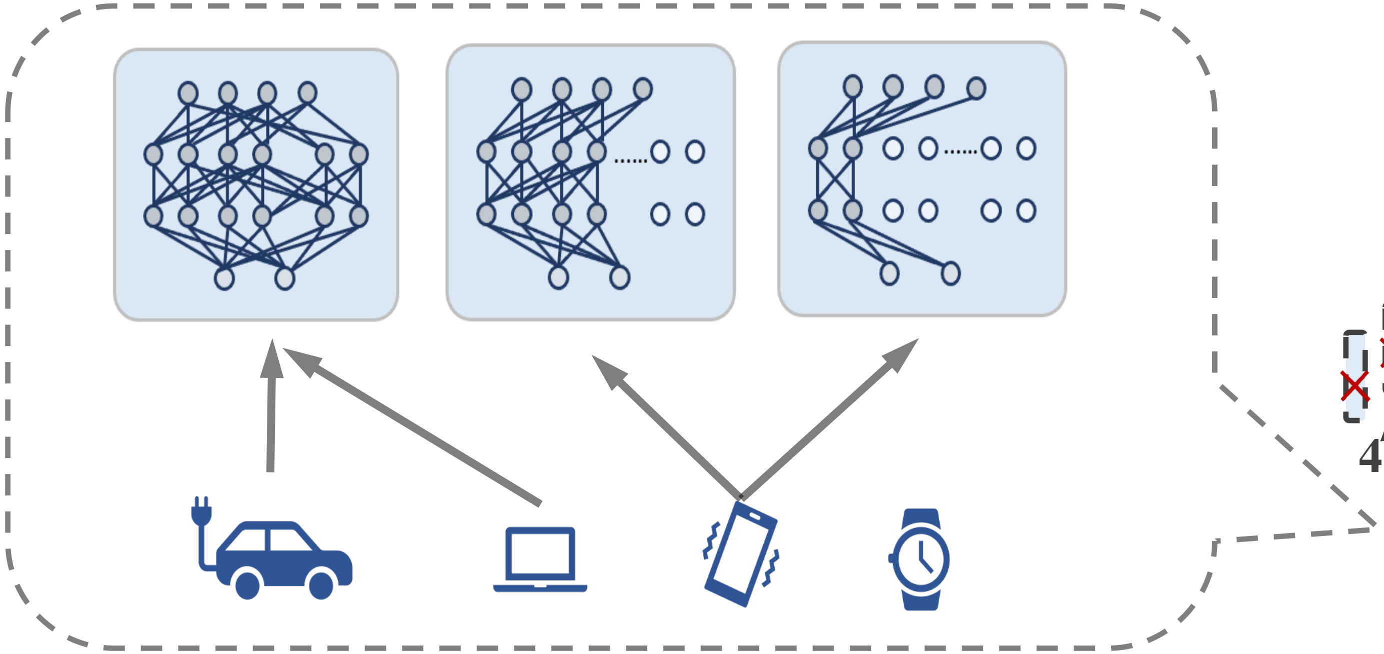
- Model parameters of columns are transmitted ***sequentially*** between the server and the client.

- The received model parameters compose a functional sub-model.



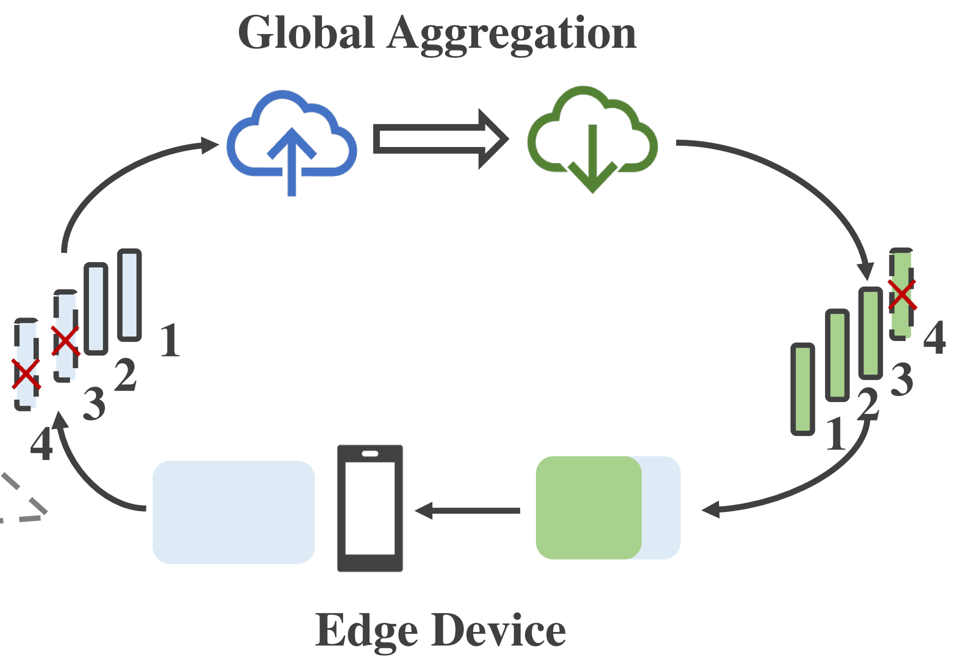
How does our approach benefit Federated Learning?

Support ***heterogeneous*** model architectures ✓



Prunable Global Model

Resilient to connection interruption ✓



Sequential Model Transmission

Paper Outline

Background and Challenges in Federated Learning

- System Heterogeneity
- Unstable connection

Motivation and Key idea

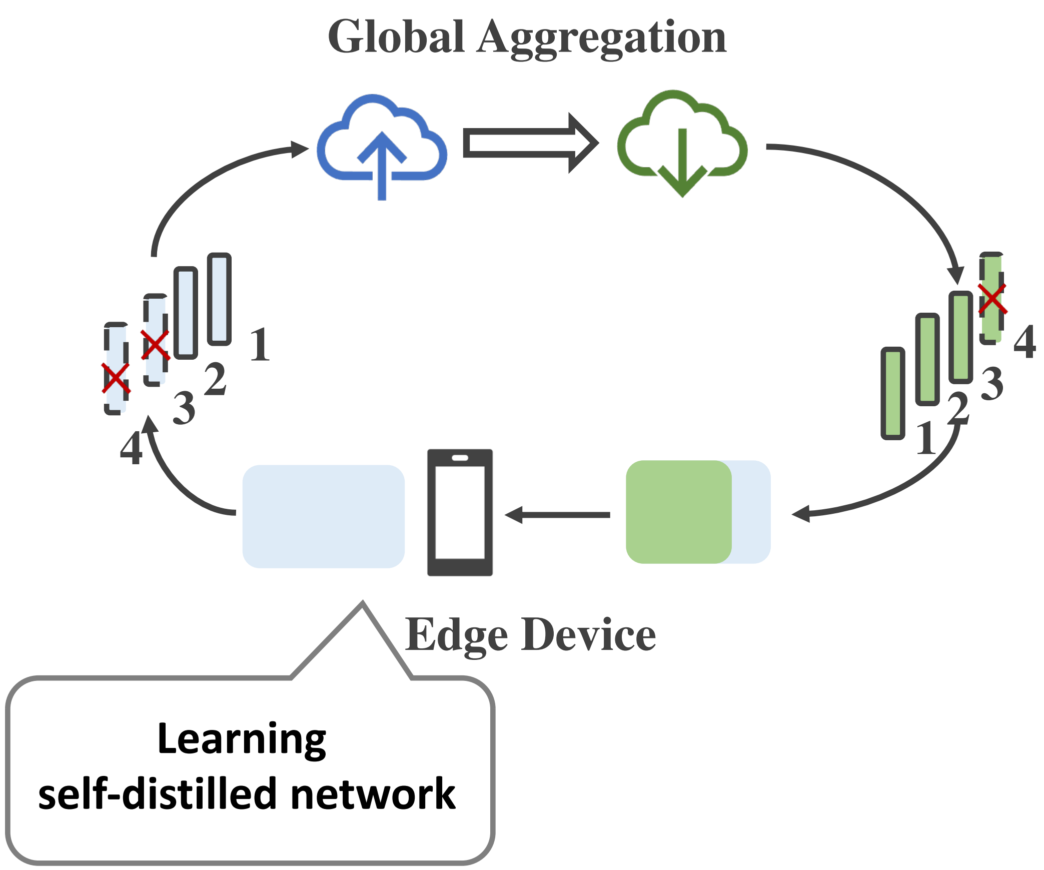
- Learning structurally prunable networks

Methodology

- Self-distilled network via progressive learning

Performance Evaluation

Proposed Approach:
Self-Distilled Network for Federated Learning

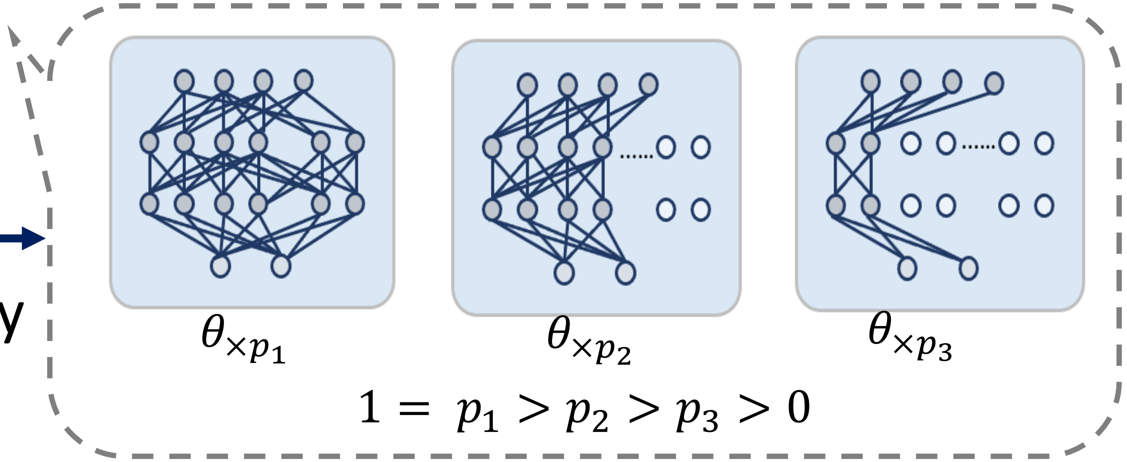


Proposed Approach: *Self-Distilled Network* for Federated Learning

Local Training Objective:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(f(\mathcal{X};\boldsymbol{\theta}),\mathcal{Y}) + \mathop{\mathbb{E}}_{p\sim\mathcal{P}}[\mathcal{L}(f(\mathcal{X};\boldsymbol{\theta}_{\times p}),\mathcal{Y})$$

Make sub-model with arbitrary pruning ratio p predictive



Proposed Approach: *Self-Distilled Network* for Federated Learning

Local Training Objective:

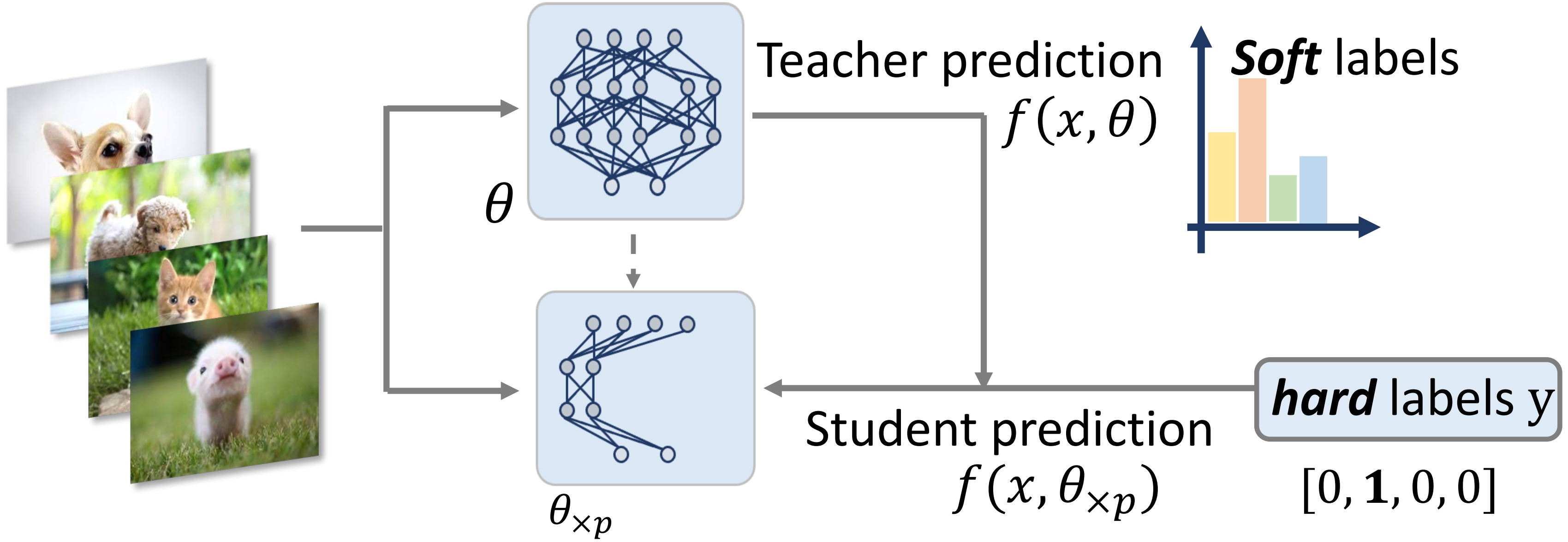
$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(f(\mathcal{X}; \boldsymbol{\theta}), \mathcal{Y}) + \mathop{\mathbb{E}}_{p \sim \mathcal{P}} [\mathcal{L}(f(\mathcal{X}; \boldsymbol{\theta}_{\times p}), \mathcal{Y})$$

- We need finer-grained guidance to assist sub-model training

Proposed Approach:
Self-Distilled Network for Federated Learning

Local Training Objective:

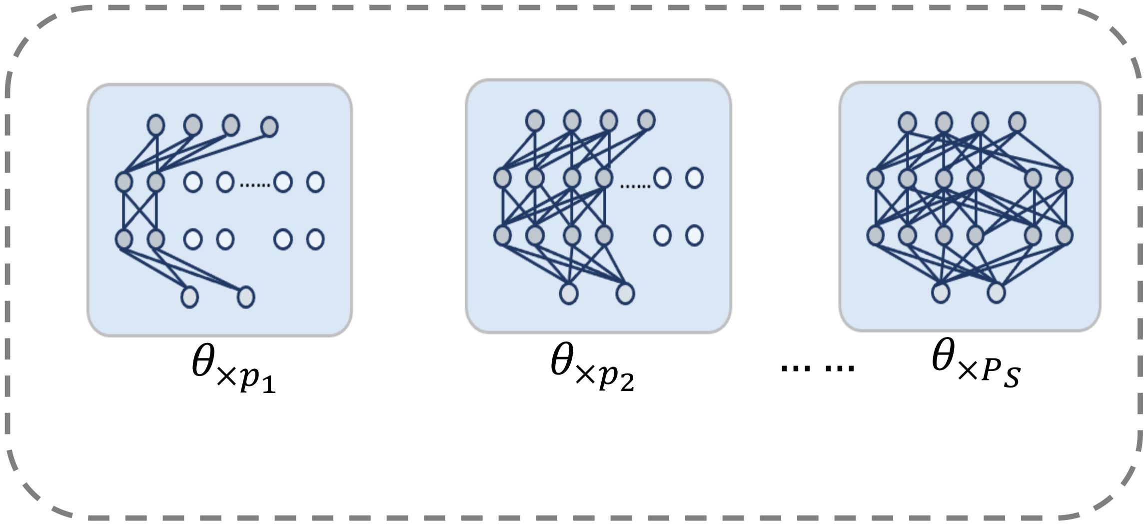
$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(f(\mathcal{X};\boldsymbol{\theta}),\mathcal{Y}) + \mathop{\mathbb{E}}_{p\sim\mathcal{P}}[\mathcal{L}(f(\mathcal{X};\boldsymbol{\theta}_{\times p}),\mathcal{Y}) + D_{\mathrm{KL}}[f(\mathcal{X};\boldsymbol{\theta})\|f(\mathcal{X};\boldsymbol{\theta}_{\times p})]], \quad (1)$$



Effective Optimization via *Progressive Learning*

1. Sample ordered pruning ratios:

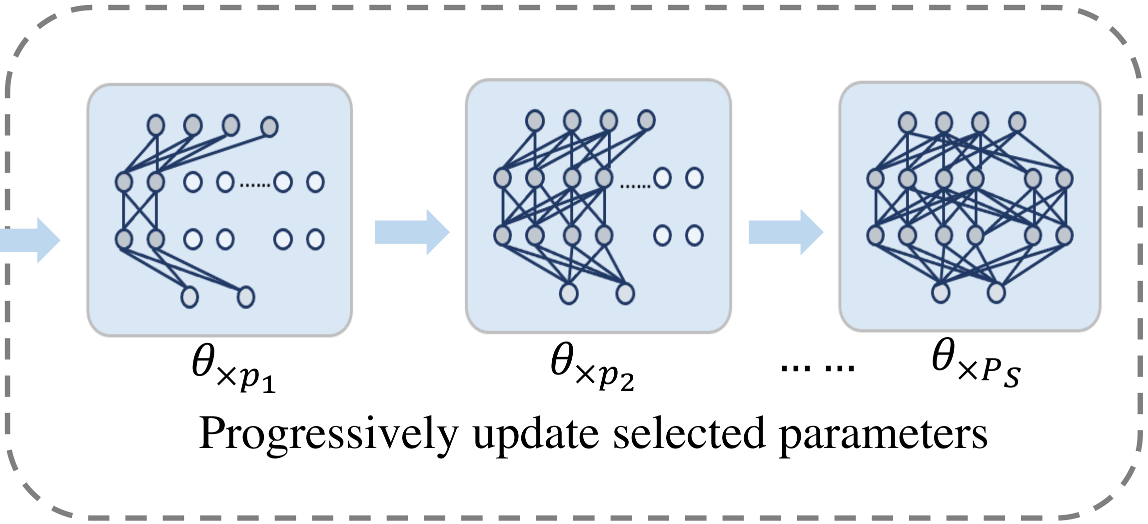
$$\hat{P} = [p_i | p_i \in P, p_i < p_{i+1}\ \forall i < S, p_S = 1.0]_{i=1}^{S}$$



Effective Optimization via *Progressive Learning*

1. Sample ordered pruning ratios:

$$\hat{P} = [p_i | p_i \in P, p_i < p_{i+1} \; \forall i < S, p_S = 1.0]_{i=1}^{S}$$



2. Progressive parameter update:

for $p_i \sim \hat{P}$ **do**

$g_i \leftarrow \nabla_{\{\boldsymbol{\theta}_{\times p_i} \backslash \boldsymbol{\theta}_{\times p_{i-1}}\}} J(x; \boldsymbol{\theta}_{\times p_i})$

$\boldsymbol{\theta}_{\times p_i} \leftarrow \boldsymbol{\theta}_{\times p_i} - \eta * g_i.$

end for

$$J(x; \boldsymbol{\theta}_{\times p_i}) = \mathcal{L}(f(x; \boldsymbol{\theta}_{\times p_i}), y) + \alpha_i D_{\mathrm{KL}}[f(x; \bar{\boldsymbol{\theta}}) \| f(x; \boldsymbol{\theta}_{\times p_i})]$$

Paper Outline

Background and Challenges in Federated Learning

- System Heterogeneity
- Unstable connection

Motivation and Key idea

- Learning structurally prunable networks

Methodology

- Self-distilled network via progressive learning

Performance Evaluation

- Robustness
- Communication Efficiency

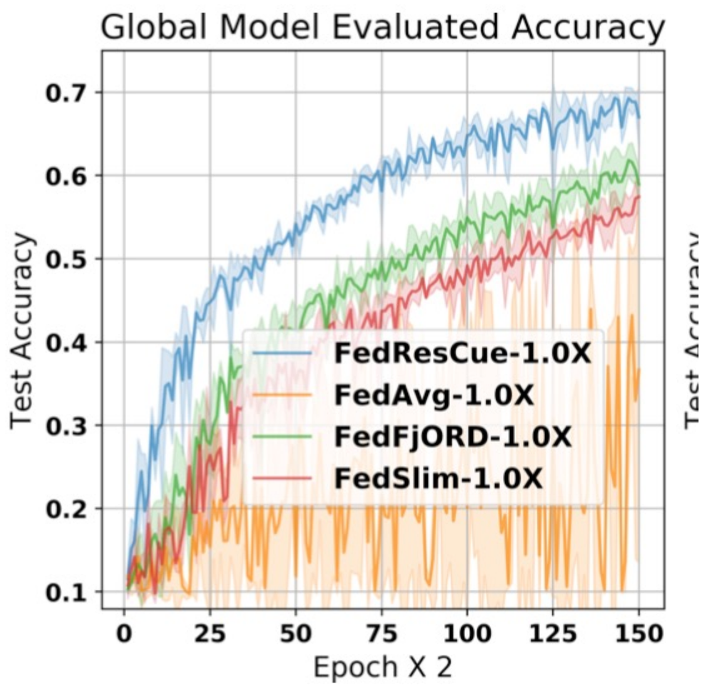
Performance Under System Heterogeneity

- Our approach:
  - consistently outperforms baselines under ***system heterogeneity*** (*i.e.* the cluster setting).
  - is more advantageous on smaller model size($\mathbf{w}_{\times 0.25}$) and fewer training data (typical scenario for FL)

Global Model Accuracy (%) Evaluated on CIFAR10, with Stable Network Connections ($er = 0$).							
Training Data	System Heterogeneity	Evaluated Model	*FedAvg*	*FedHetero*	*FjORD*	*FedSlim*	*FedResCuE*
100%	$P_c = \{1.0\}$ (uniform)	$\boldsymbol{w}_{\times 1}$	81.06±0.63	-	80.57±0.91	81.14±0.76	**81.39±0.20**
		$\boldsymbol{w}_{\times 0.25}$	18.57±0.64	-	69.94±0.65	70.47±0.61	**71.19±0.19**
	$\lvert P_c\rvert = 4$ (cluster)	$\boldsymbol{w}_{\times 1}$	-	76.80±0.53	75.71±0.47	77.49±0.40	**78.22±0.41**
		$\boldsymbol{w}_{\times 0.25}$	-	68.56±0.51	70.98±0.75	73.22±0.34	**73.25±0.47**
20%	$P_c = \{1.0\}$ (uniform)	$\boldsymbol{w}_{\times 1}$	68.03±0.50	-	67.89±1.47	67.96±0.72	**71.27±0.27**
		$\boldsymbol{w}_{\times 0.25}$	16.47±2.24	-	**61.38±1.69**	59.56±1.39	61.12±1.35
	$\lvert P_c\rvert = 4$ (cluster)	$\boldsymbol{w}_{\times 1}$	-	59.38±0.41	62.43±1.65	59.53±0.86	**64.53±1.06**
		$\boldsymbol{w}_{\times 0.25}$	-	55.41±0.39	61.86±1.21	58.31±0.23	**61.98±0.85**

Performance Under Unstable Connections

- Our approach is more resilient to transmission package loss compared with other approaches that are compatible with system heterogeneity.



Learning curves evaluated on the ×1.0 model.

Global Model Accuracy (%) Evaluated on CIFAR10 Under Connection Loss ($er > 0$).						
System Heterogeneity	Evaluated Model	*FedAvg*	*FedHetero*	*FjORD*	*FedSlim*	***FedResCuE***
$P_c = \{1.0\}$ (uniform)	$\boldsymbol{w}_{\times 1}$	50.36±2.17	-	61.79±1.62	57.31±1.27	**70.02±0.40**
	$\boldsymbol{w}_{\times 0.25}$	12.58±0.51	-	60.20±1.67	55.33±0.89	**67.40±0.84**
$\lvert P_c \rvert = 4$ (cluster)	$\boldsymbol{w}_{\times 1}$	-	60.92±1.33	64.52±0.60	62.35±1.76	**69.78±0.74**
	$\boldsymbol{w}_{\times 0.25}$	-	59.70±0.64	64.11±0.41	61.77±1.62	**68.83±1.00**

Performance under unstable network connections, given 100% of training data, and $0.1 \le er \le 0.2$

Communication Efficiency

- Our approach requires fewer communication rounds to reach pre-defined performance.

Communication Efficiency on CIFAR10 dataset.					
Acc	Model Size	*FedHetero*	*FjORD*	*FedSlim*	*FedResCuE*
	100 % training data, $0.1 \leq er \leq 0.2$.				
60%	$\boldsymbol{w}_{\times 0.5}$	256.7	218.0	253.3	**124.7**
	20 % training data, $er = 0$				
55%	$\boldsymbol{w}_{\times 0.5}$	180.7	156.0	192.0	**96.0**

Table 6: *FedResCuE* requires notably fewer communication rounds to reach the predefined accuracy (Acc).

Contribution Overview

FedResCue

- Self-Distillation
- System Heterogeneity
- Structurally Prunable Network
- Communication Efficiency
- Robustness to Connection Uncertainty