FedMask : Joint Computation and Communication-Efficient Personalized Federated Learning via Heterogeneous Masking

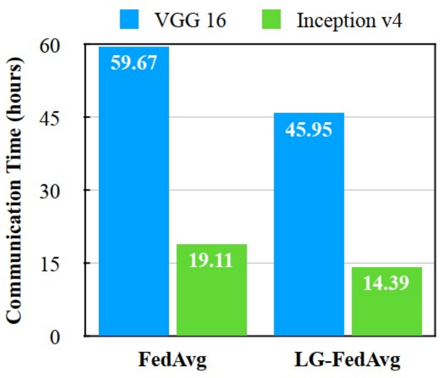
Du Xiao - March, 11th, 2022

Chanllege of federated learning

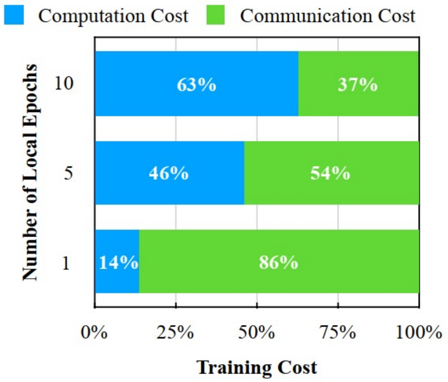
Communication Bottleneck

Computation Contraint

Need Personalization for Heterogenous



Comparison of communication cost between FedAvg and LG-FedAvg.



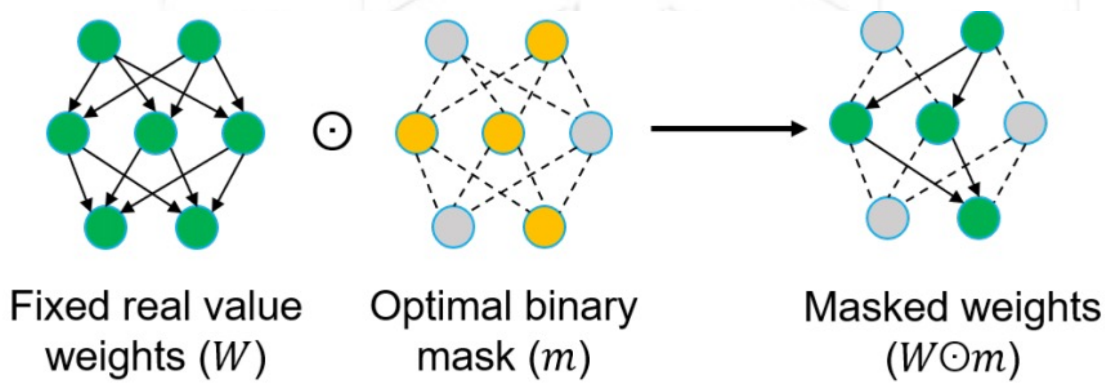
Training cost with different numbers of local training epochs.

Motivate

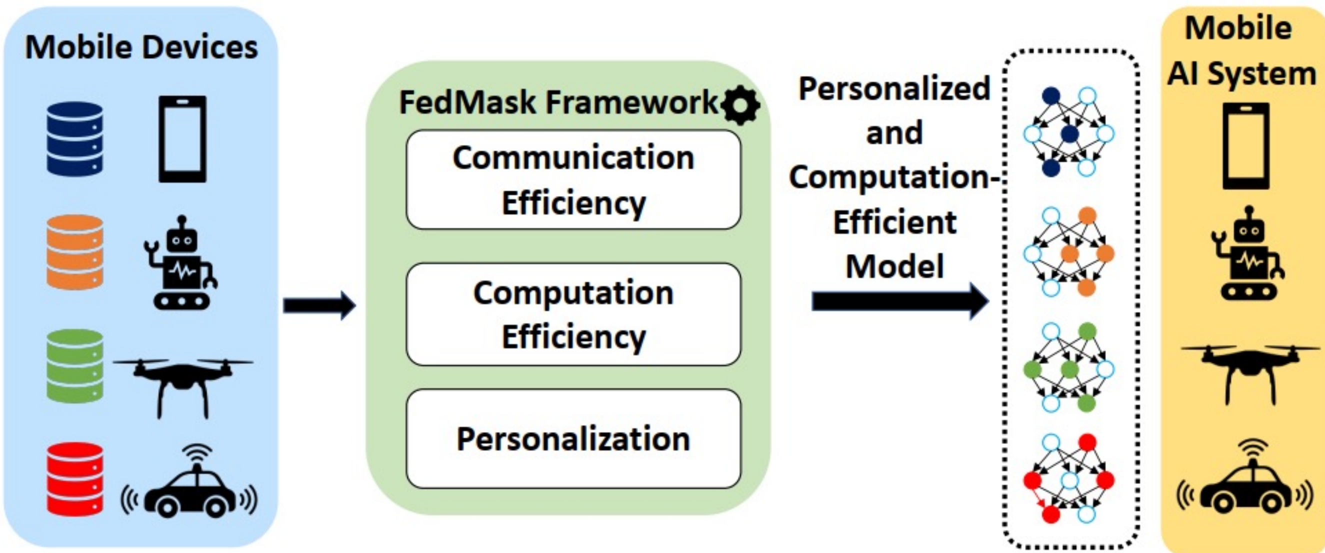
Minimize the communication cost

Reduce computation cost for training

Learn a personalized model for each device to mitigate statistical heterogeneity

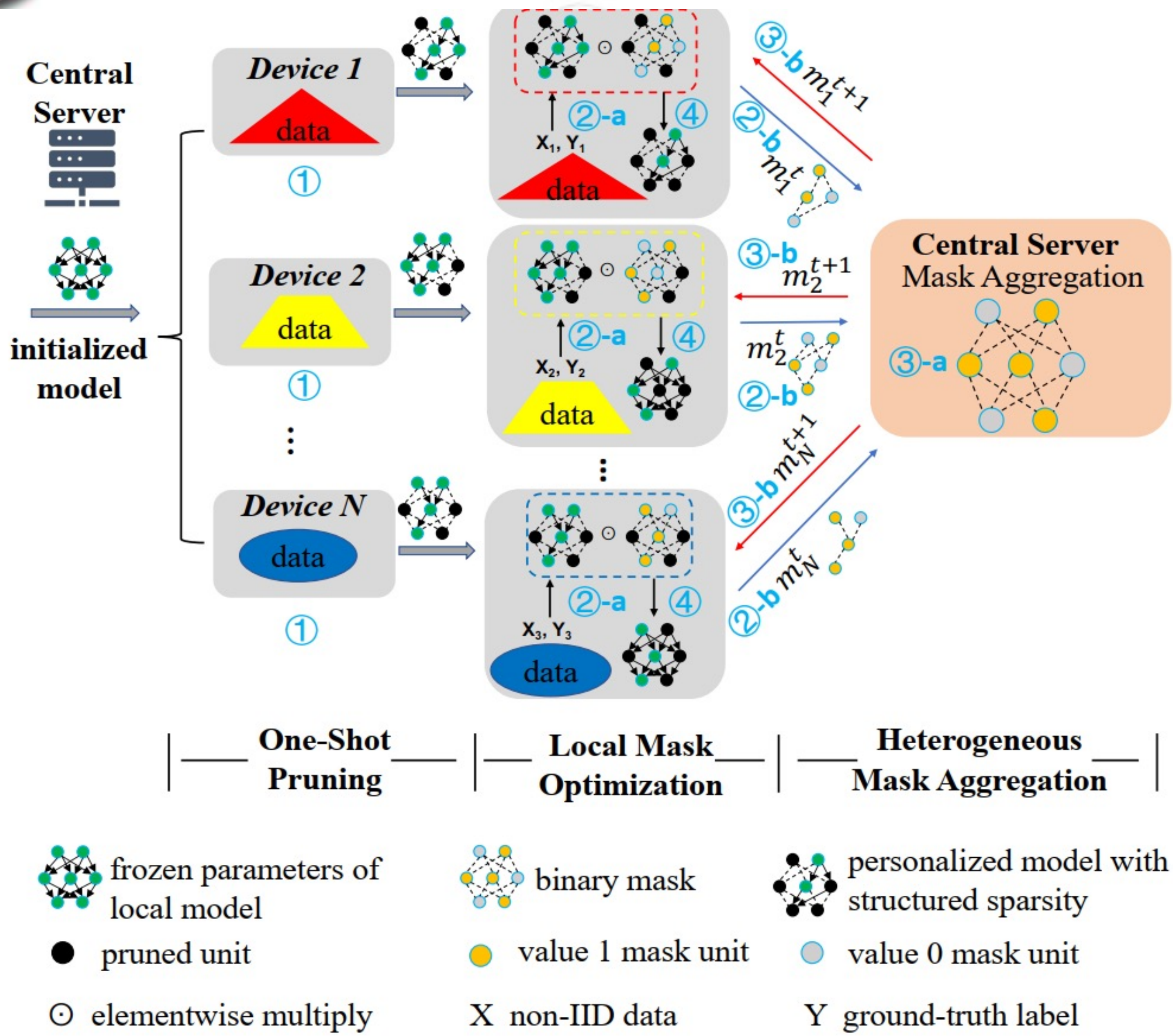


Motivate



Method	Computation Efficiency	Communication Efficiency	Personalization
FedAvg [42]	X	X	X
Top-k [1]	X	✓	X
Per-FedAvg [13]	X	X	✓
LG-FedAvg [37]	X	✓	✓
FedMask	**✓**	**✓**	**✓**

Overview of FedMask



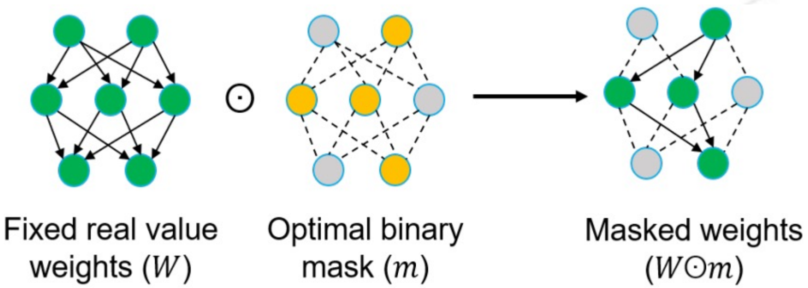
Design Chanlleges

How to jointly improve communication and computation efficiency

How to incorporate personalization for each device?

How to aggregate heterogeneous binary masks on the central server while preserving personalization?

Binary Mask Optimization



$$y = (W \odot m) \cdot x,$$

we introduce a realvalued mask m_r:

$$m_{ij} = \begin{cases} 1, & m^r_{ij} \geq \tau \\ 0, & m^r_{ij} < \tau \end{cases} \quad \frac{\partial L}{\partial m} = \left(\frac{\partial L}{\partial y} \cdot x^T \right) \odot W,$$

$$m_{ij} = \sigma\left(m^r_{ij}\right).$$

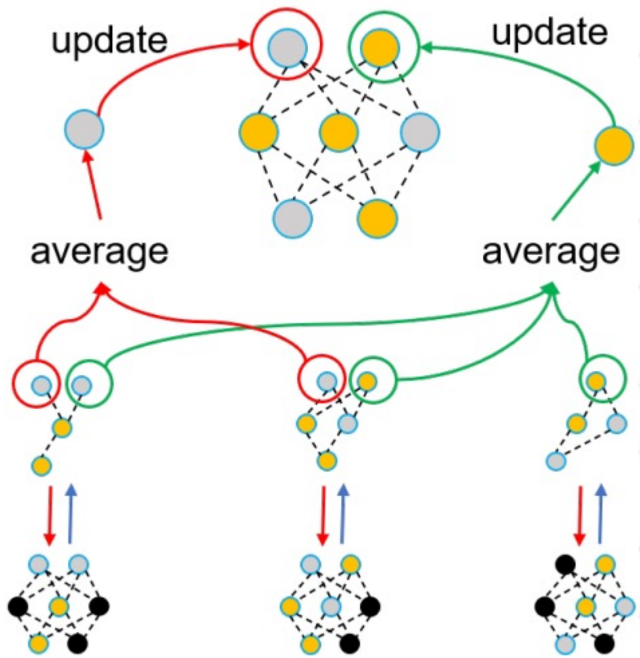
$\sigma(\cdot)$ is differentiable sigmoid function

existing optimization algorithms to m due to its binary value

One-Shot Pruning for Mask

- Pruing the mask based on $W \odot m^r$
- Preserves the dense structure of the top layers in the binary mask and only prunes the last serveral layers which compose the classifier part
- Initialize a heterogenous mask for each device

Aggregate Heterogeneous Binary Masks



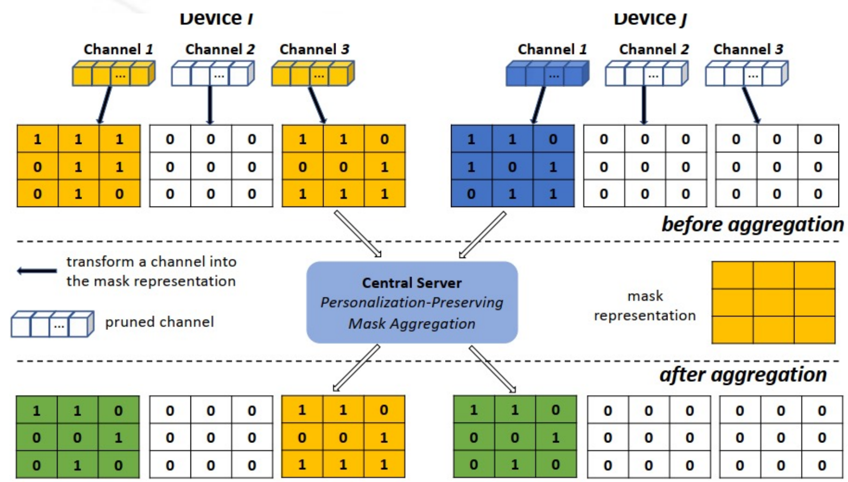


Figure 8: Illustration of the personalization-preserving mask aggregation on the central server. The *yellow* and *blue* matrices represent the unpruned masks, the green ones stand for the updated masks which are intersected between devices.

personalization

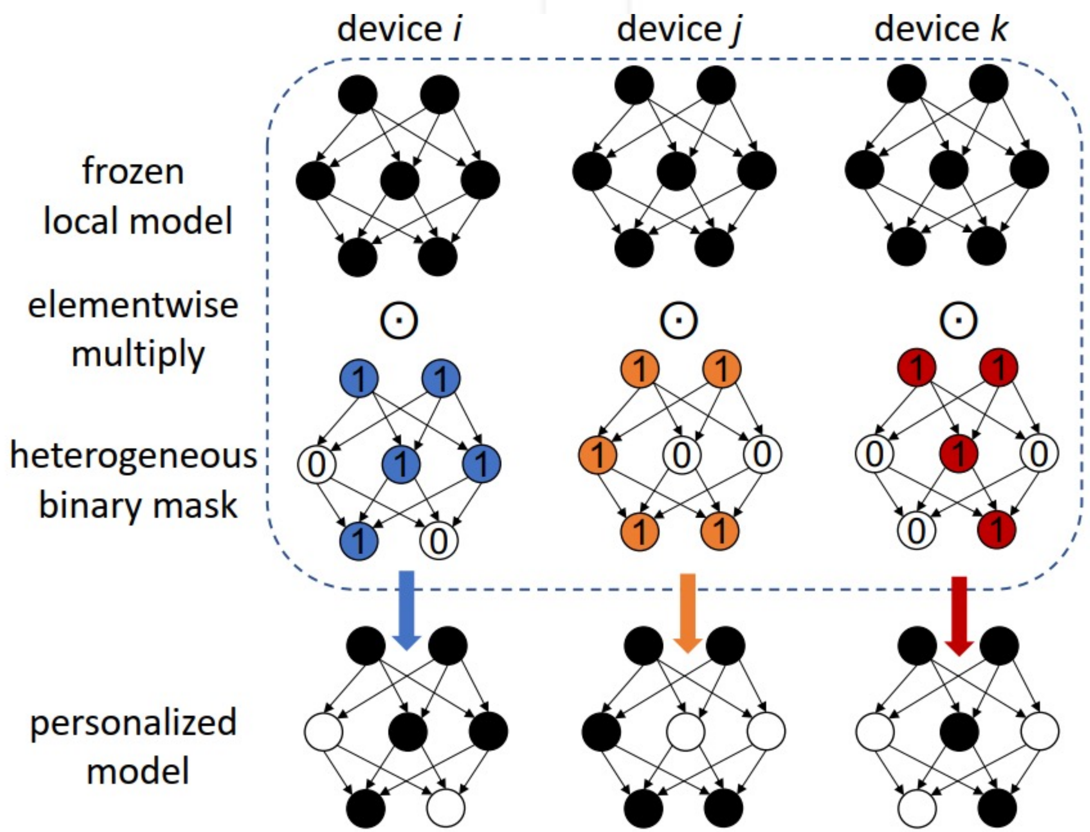


Figure 9: Illustration of achieving personalization via heterogeneous binary masks.

Evaluation

Dataset	Number of devices	Average samples per device	Classes	Non-IID
EMNIST [8]	2414	232.8	64	✓
CIFAR10 [31]	400	25	10	✓
HAR [3]	30	364.3	6	✓
Shakespeare [42]	1129	3743.2	80	✓

Evaluation

Compare with 6 baselines:

- Standalone, FedAvg,Top-k,BNN-FedAvg,Per-FedAvg,LG-FedAvg

Evaluation Metrics:

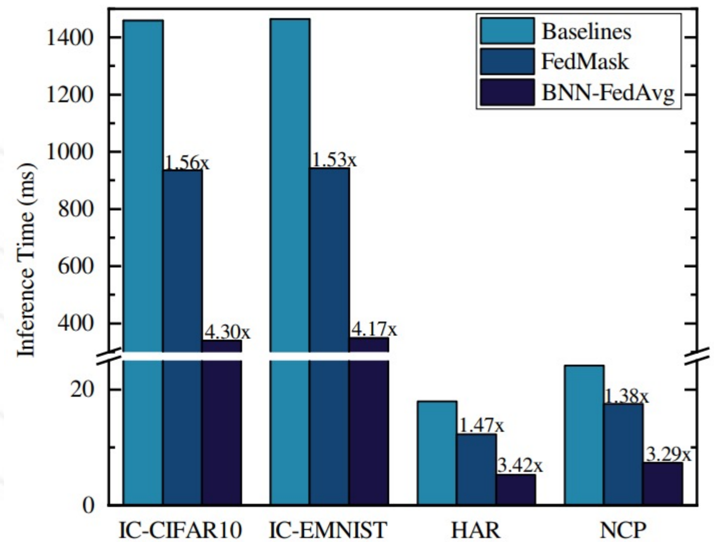
- Training Performance:
  Inference accuracy, communication cost, computation cost

- Runtime Performance:
  Memory footprint, inference time, energy consumption

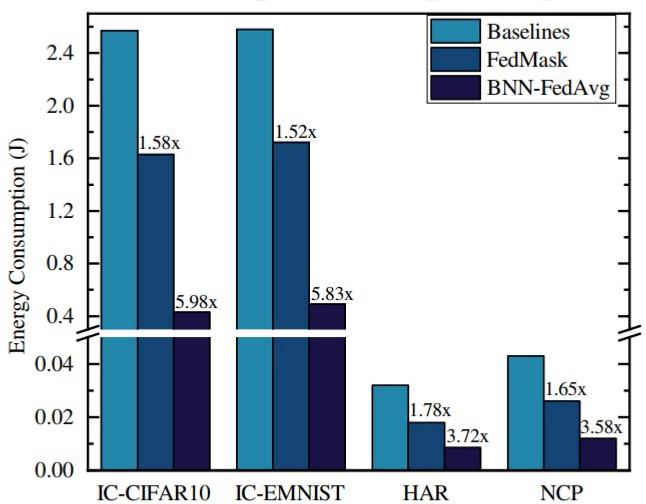
Runtime Performance

Table 5: Memory footprint reduction of FedMask.

Application	FedMask Model Size (MB)	Baseline Model Size (MB)	BNN-FedAvg Model Size (MB)
IC-CIFAR10	365.30	537.21	16.78
IC-EMNIST	364.72	538.09	16.82
HAR	2.69	4.41	0.14
NCP	0.92	1.53	0.05
All Included	733.63	1081.24	33.79



Inference speed



Reduction on energy consumption

Training Performance:

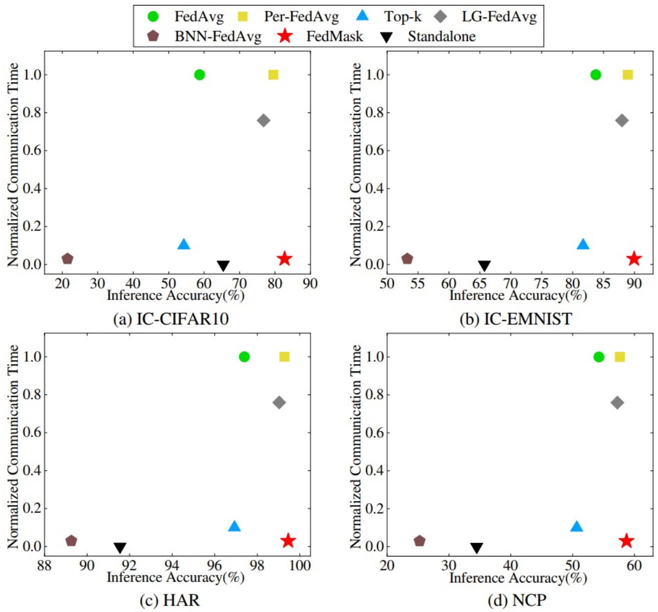


Figure 10: Comparison between FedMask and baselines in inference accuracy-communication cost space.

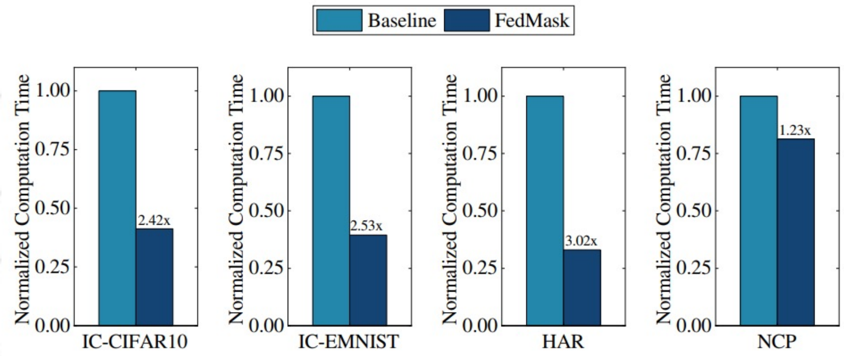


Figure 11: Comparison between FedMask and baselines in computation cost space.

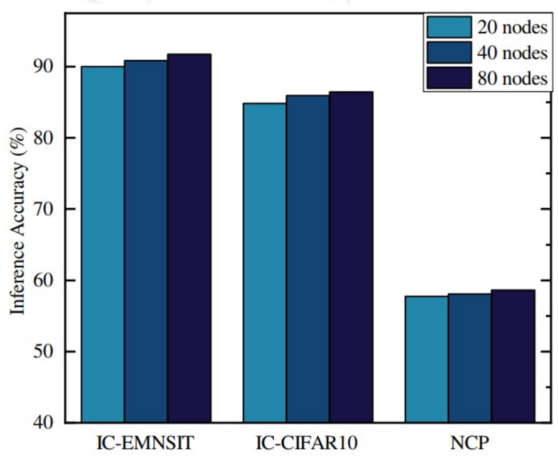


Figure 12: The impact of the number of participating devices on FedMask performance.