UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

ANTS LAB

Parameterized Knowledge Transfer for Personalized Federated Learning

Jie Zhang, et al

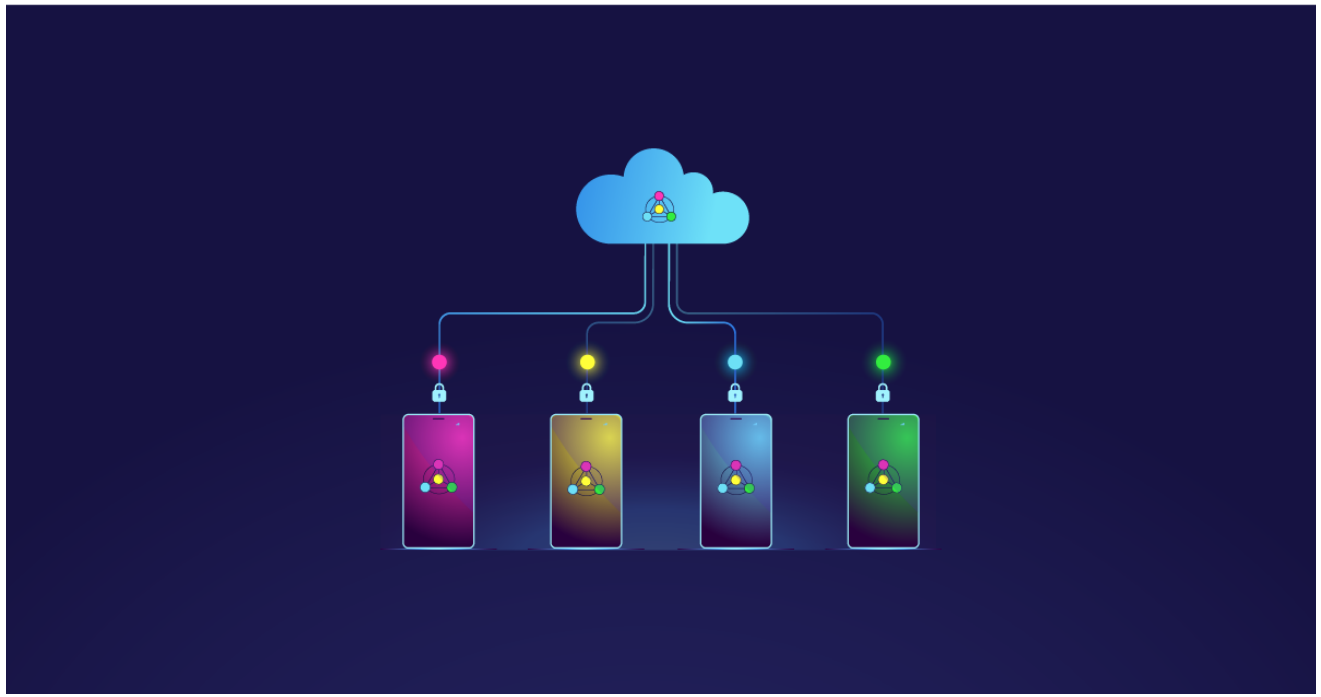
Published in NIPS' 21

Presenter: Dian Shi

Outline

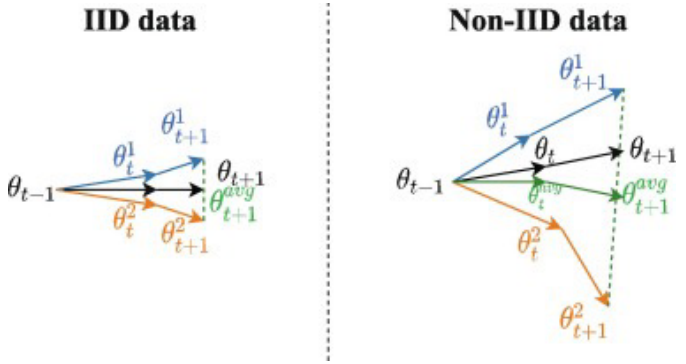
- Motivations of personalized FL (pFL)
- Knowledge Distillation personalized FL (KT-pFL)
- Performance Evaluation
- Conclusion

Federated Learning



- **Challenge:**

  Data heterogeneity (non-IID)



Global Model: cannot be generalized well

↓

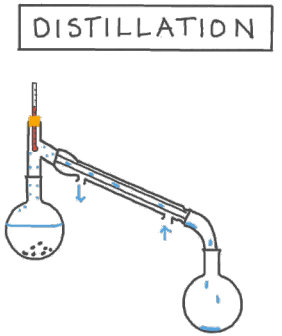
Personalized Federated Learning

Personalized Federated Learning

Personalize the global model for each client

- Existing approaches of pFL:
  regularization-based methods or meta-learning-based methods

  Only consider the homogenous model settings, impractical



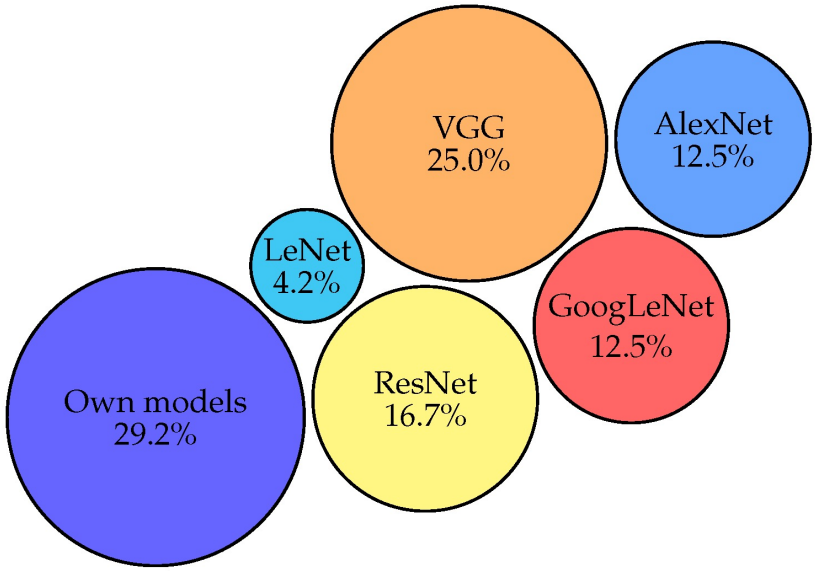
- Knowledge Distillation (KD) :
  Knowledge can be transferred among heterogenous models

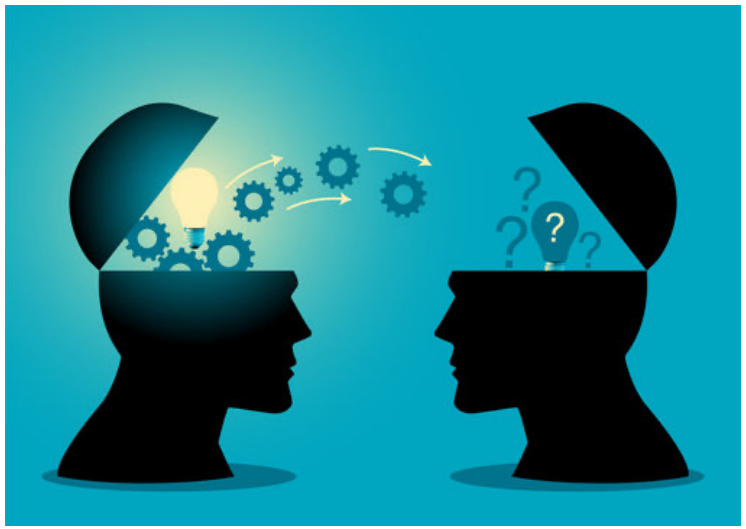
  Ignore the personalized requirements

Personalized Federated Learning

Goal: Personalized Federated Learning

- accommodate heterogeneous model structures
- achieve personalized knowledge transfer





Outline

- Motivations of personalized FL (pFL)
- Knowledge Distillation personalized FL (KT-pFL)
- Performance Evaluation
- Conclusion

Problem Formulation

Collaboratively train personalized models

N clients: $\mathbb{D}_n := \{x_i^n, y_i\}$ $x_i, y_i \in \{1, 2, \cdots, C\}$

Conventional FL:

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) := \sum_{n=1}^{N} \frac{D_n}{D} \mathcal{L}_n(\mathbf{w}), \text{ where } \mathcal{L}_n(\mathbf{w}) = \frac{1}{D_n} \sum_{i=1}^{D_n} \mathcal{L}_{CE}(\mathbf{w}; x_i, y_i)$$

- Require to have a unified model structure → Heterogenous model
- Minimize the total empirical loss → Heterogenous data

Problem Formulation

Personalized loss function of client n:

collaborative knowledge

Data sample from public dataset

knowledge coefficient

$$\mathcal{L}_{per,n}(\mathbf{w}^n) := \mathcal{L}_n(\mathbf{w}^n) + \lambda \sum_{\hat{x} \in \mathbb{D}_r} \mathcal{L}_{KL}\left(\sum_{m=1}^{N} c_{mn} \cdot s(\mathbf{w}^m, \hat{x}), s(\mathbf{w}^n, \hat{x}) \right)$$

$$\mathcal{L}_n(\mathbf{w}) = \frac{1}{D_n} \sum_{i=1}^{D_n} \mathcal{L}_{CE}(\mathbf{w}; x_i, y_i)$$

Kullback–Leibler (KL) Divergence function

- *KL loss function*: transfer personalized knowledge from a teacher to another
- c_{mn}: estimate the contribution from client m to n

Problem Formulation

Personalized loss function of client n:

collaborative knowledge

$$\mathcal{L}_{per,n}(\mathbf{w}^n) := \mathcal{L}_n(\mathbf{w}^n) + \lambda \sum_{\hat{x} \in \mathbb{D}_r} \mathcal{L}_{KL}\left(\sum_{m=1}^{N} c_{mn} \cdot s(\mathbf{w}^m, \hat{x}), s(\mathbf{w}^n, \hat{x}) \right)$$

teacher

student

collaborative knowledge : soft predictions or model parameters

The output of the last fully connected layer

$$\text{i.e., } s(\mathbf{w}^n, \hat{x}) = \frac{\exp(z_c^n / T)}{\sum_{c=1}^{C} \exp(z_c^n / T)}$$

Temperature hyperparameter of the softmax function

Problem Formulation

Personalized loss function of client n:

knowledge coefficient

$$\mathcal{L}_{per,n}(\mathbf{w}^n) := \mathcal{L}_n(\mathbf{w}^n) + \lambda \sum_{\hat{x} \in \mathbb{D}_r} \mathcal{L}_{KL}\left(\sum_{m=1}^{N} c_{mn} \cdot s(\mathbf{w}^m, \hat{x}), s(\mathbf{w}^n, \hat{x}) \right)$$

knowledge coefficient matrix: $\mathbf{c} \in \mathbb{R}^{N \times N}$

$$\mathbf{c} = \begin{Bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{Bmatrix}$$

Our objective:

Trade-off personalization and generalization

identity matrix

$$\min_{\mathbf{w},\mathbf{c}} \mathcal{L}(\mathbf{w}, \mathbf{c}) := \sum_{n=1}^{N} \frac{D_n}{D} \mathcal{L}_{per,n}(\mathbf{w}^n) + \rho \|\mathbf{c} - \frac{\mathbf{1}}{N}\|^2$$

regularization term

$\mathbf{w} = [\mathbf{w}^1, \cdots, \mathbf{w}^N]$

Knowledge Transfer-Personalized FL (KT-pFL)

Our objective:

$$\min_{\mathbf{w},\mathbf{c}} \mathcal{L}(\mathbf{w},\mathbf{c}) := \sum_{n=1}^{N} \frac{D_n}{D} \mathcal{L}_{per,n}(\mathbf{w}^n) + \rho \|\mathbf{c} - \frac{\mathbf{1}}{N}\|^2$$

$$\mathcal{L}_{per,n}(\mathbf{w}^n) := \mathcal{L}_n(\mathbf{w}^n) + \lambda \sum_{\hat{x} \in \mathbb{D}_r} \mathcal{L}_{KL} \left(\sum_{m=1}^{N} c_{mn} \cdot s(\mathbf{w}^m, \hat{x}), s(\mathbf{w}^n, \hat{x}) \right)$$

Solutions: KT-pFL

Main idea: the local model parameters (w) and knowledge coefficient matrix (c) are updated alternatively

Knowledge Transfer-Personalized FL (KT-pFL)

Solutions: KT-pFL

$$\mathcal{L}_{per,n}(\mathbf{w}^n) := \mathcal{L}_n(\mathbf{w}^n) + \lambda \sum_{\hat{x}\in\mathbb{D}_r} \mathcal{L}_{KL}\left(\sum_{m=1}^{N} c_{mn}\cdot s(\mathbf{w}^m,\hat{x}), s(\mathbf{w}^n,\hat{x})\right)$$

$$\min_{\mathbf{w},\mathbf{c}} \mathcal{L}(\mathbf{w},\mathbf{c}) := \sum_{n=1}^{N} \frac{D_n}{D}\mathcal{L}_{per,n}(\mathbf{w}^n) + \rho\|\mathbf{c} - \frac{\mathbf{1}}{N}\|^2$$

Update **w**: locally, two-stage framework

- Local training: Train w on clients' private data

$$\mathbf{w}^n \leftarrow \mathbf{w}^n - \eta_1 \nabla_{\mathbf{w}^n}\mathcal{L}_n(\mathbf{w}^n;\xi_n)$$

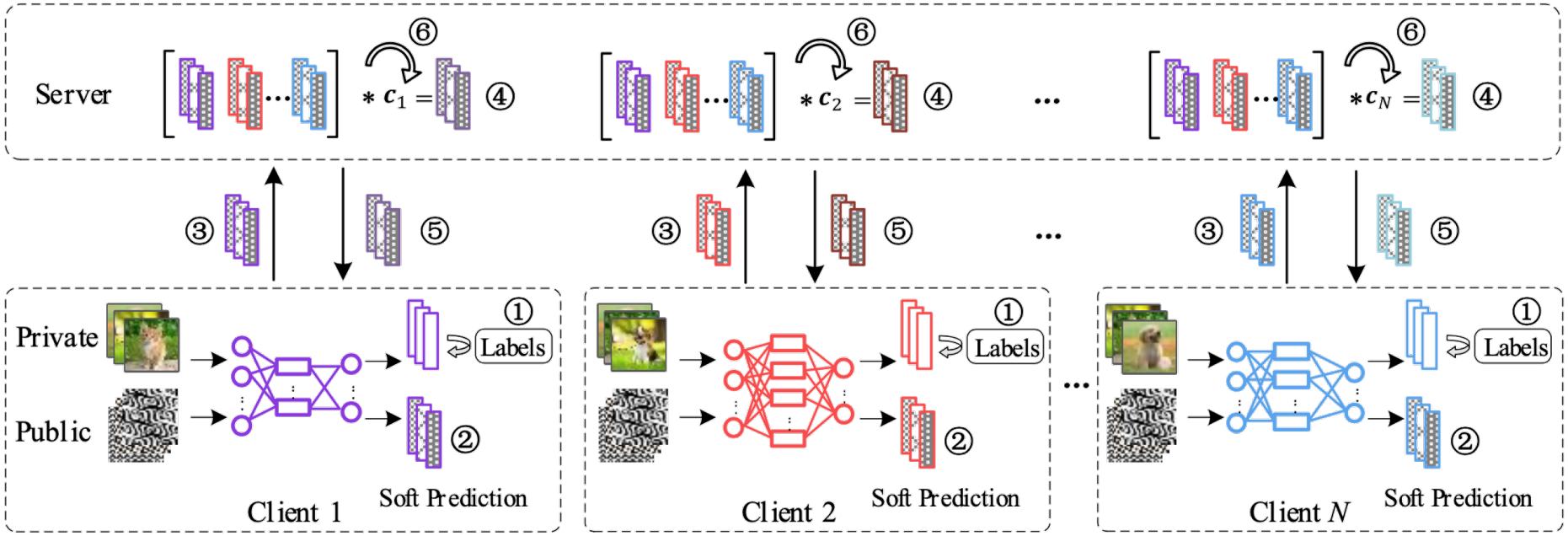
- distillation: Transfer knowledge from personalized soft prediction based on public data

$$\mathbf{w}^n \leftarrow \mathbf{w}^n - \eta_2 \nabla_{\mathbf{w}^n}\mathcal{L}_{KL}\left(\sum_{m=1}^{N} \mathbf{c}_m^{*,T}\cdot s(\mathbf{w}^m,\xi_r), s(\mathbf{w}^n,\xi_r)\right)$$

Update **c**: update c in the server

$$\mathbf{c} \leftarrow \mathbf{c} - \eta_3\lambda\sum_{n=1}^{N}\frac{D_n}{D}\nabla_{\mathbf{c}}\mathcal{L}_{KL}\left(\sum_{m=1}^{N}\mathbf{c}_m\cdot s(\mathbf{w}^{m,*},\xi_r), s(\mathbf{w}^{n,*},\xi_r)\right) - 2\eta_3\rho(\mathbf{c}-\frac{\mathbf{1}}{N}),$$

Knowledge Transfer-Personalized FL (KT-pFL)



① Local training on private data

② Output the local soft prediction on public data

③ Send the soft prediction to server

④ Calculate each client's personalized soft prediction

⑤ Download the personalized soft prediction to perform distillation phase

⑥ Update the knowledge coefficient matrix

Outline

- Motivations of personalized FL (pFL)
- Knowledge Distillation personalized FL (KT-pFL)
- Performance Evaluation
- Conclusion

Experimental Settings (KT-pFL)

Private datasets: EMNIST, Fashion_MNIST, and CIFAR-10

Public datasets: MNIST CIFAR-100

Model Structures: LeNet, AlexNet, ResNet-18, ShuffleNet,
20 clients: five clients per model

Learning settings: 20 local training epochs and 1 distillation step

Baselines:

non-personalized distillation-based methods

FedMD: utilize the aggregated softmax scores to guide the local model

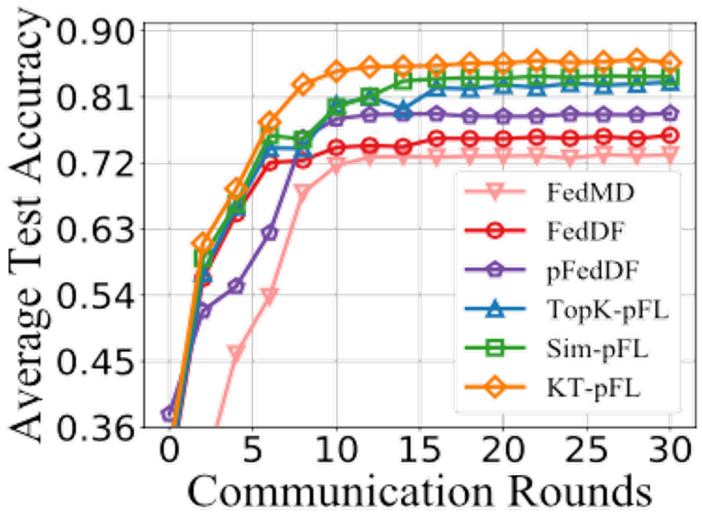
FedDF: only maintain the global model

personalized distillation-based methods

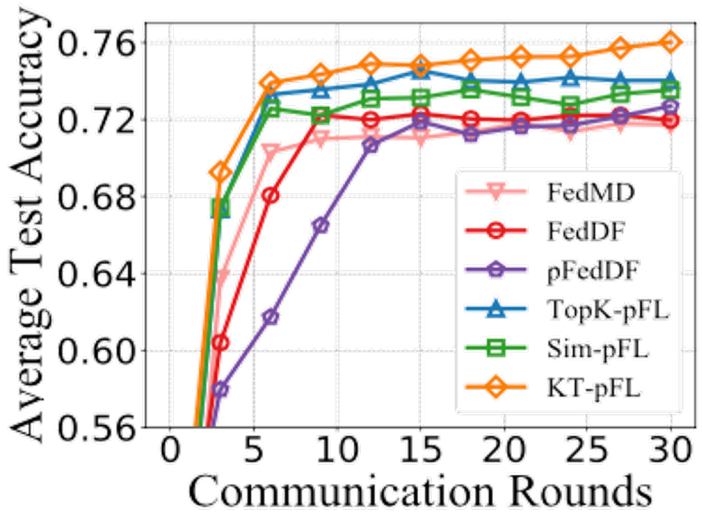
pFedDF: fine-tune the FedDF with own dataset Sim-pFL TopK-pFL

Results

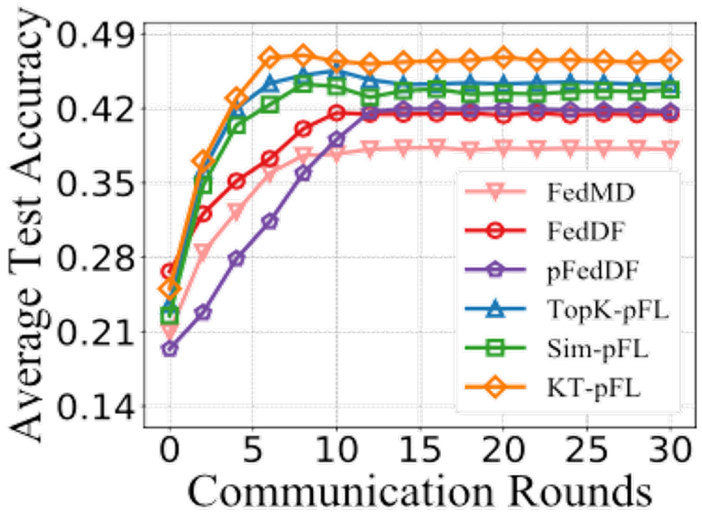
Non-IID case: each client contains all labels, while the number of samples for each class is different from that of a different client.



(a) EMNIST



(b) Fashion_MNIST



(c) CIFAR-10

- KT-pFL obtains better accuracy performance than others.
- The performances of FedMD, FedDF, pFedDF are worse than the others. (produce only one global soft prediction, cannot be adapted to each client)

Some Observations

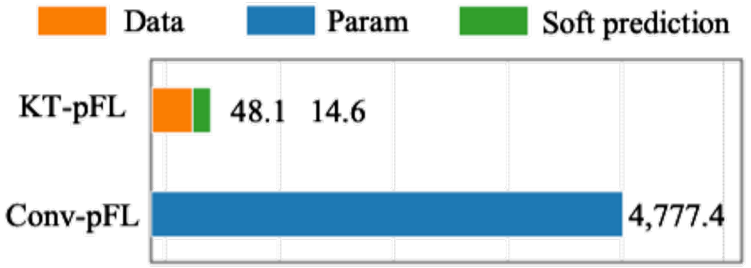
- Effect of Local Epochs *E*: trade-off between the computations and communications, i.e., larger E requires more comp. at local clients, smaller E needs more global comm.
- Effect of Distillation Steps *R*: larger value of R cannot lead to better performance

Dataset	# Local epochs (E)				# Distillation steps (R)			
	5	10	15	20	1	2	3	5
EMNIST (%)	90.15	**92.76**	91.03	90.15	**91.76**	91.40	91.18	91.54
Fashion_MNIST (%)	88.73	89.14	88.58	**89.42**	89.14	**89.90**	89.07	88.08
CIFAR-10 (%)	58.30	**59.38**	59.34	59.24	**59.24**	57.99	59.22	58.91

- Effect of public dataset: different public datasets have little effect on the performance.

Public dataset	Test Accuracy
MNIST	89.14 ($\pm$ 0.15)
EMNIST	88.76 ($\pm$ 0.21)
Unlabeled Open Dataset	89.03 ($\pm$ 0.11)

- Efficiency Evaluation: communication overhead



Conclusion

- This paper is the first to study the personalized knowledge transfer in FL.
- Propose the 'knowledge coefficient matrix' to identify the contribution from one client to others' local training.
- KT-pFL outperforms the existing approaches in terms of the final accuracy and communication overhead.


THANK YOU
UNIVERSITY of HOUSTON | ENGINEERING