Knowledge-Enhanced Semi-Supervised Federated Learning for Aggregating Heterogeneous Lightweight Clients in IoT

Authors: Jiaqi Wang[1*], Shenglai Zeng[2*], Zewei Long[3], Yaqing Wang[4], Houping Xiao[5], Fenglong Ma[1]
Presentor: Jiaqi Wang
3:15 PM - 5:15 PM, Friday, 04/28/2023
Room Think 3

*equal contribution, [1]The Pennsylvania State University, [2]University of Electronic Science and Technology of China, [3]University of Illinois Urbana-Champaign, [4]Purdue University, [5]Georgia State University

Content

1/ Background

2/ Motivation

3/ Proposed Work

4/ Experiment Results

5/ Conclusion

Content

1/ Background

2/ Motivation

3/ Proposed Work

4/ Experiment Results

5/ Conclusion

Background

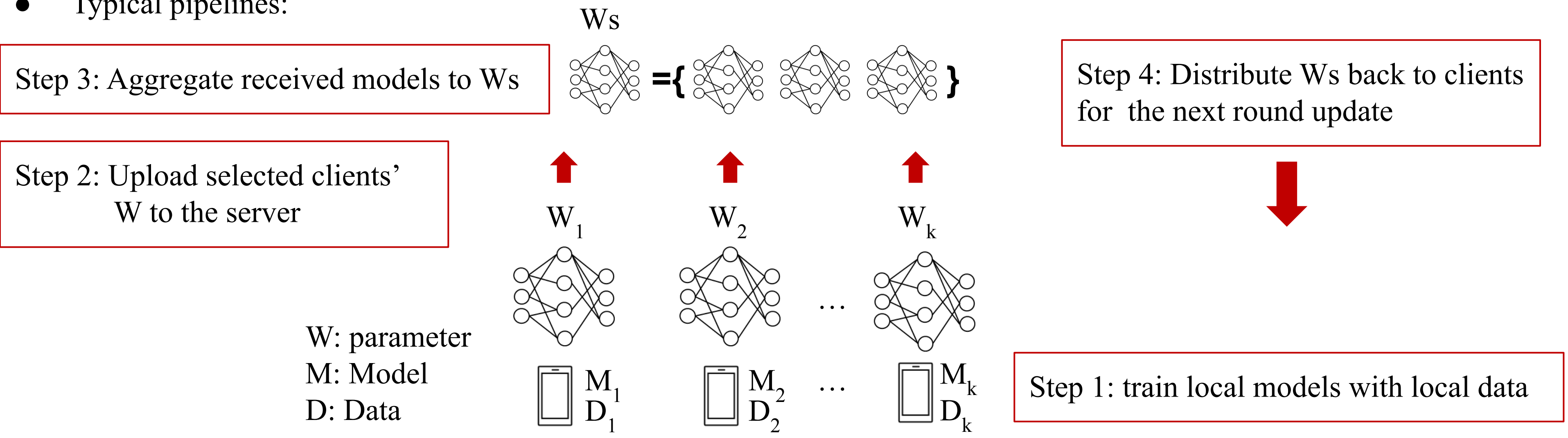
- Data grows tremendously with more and more data acquisition techniques, e.g., Google processes 8.5 billion searches[1], WhatsApp users exchange up to 65 billion messages daily[2], the number of IoT devices could rise to 41.6 billion by 2025[3].
- There are regulations and laws being created to protect the data acquisition, data use, and data share, such as The American Data Privacy Protection Act (ADPPA), General Data Protection Regulation (GDPR), and California Privacy Rights Act (CPRA).
- However, more and more data privacy concerns are raised, e.g., 72% of US adults think that companies have too much control over their personal data[1], which causes public privacy concern and social trustworthiness[5].

[1] Data source: Oberlo
[2] Data source: Connectiva Systems
[3] Source: IDC
[4] Source: YouGov
[5] Wang, Jiaqi. "An In-depth Review of Privacy Concerns Raised by the COVID-19 Pandemic." arXiv preprint arXiv:2101.10868 (2021).

How can we utilize the data saved distributedly and
enable different data holders to cooperate with the data privacy guarantee?

Federated Learning

- Federated learning (FL) [1], as one of the solutions, enables multiple clients to train models collaboratively by only sharing model parameters, instead of sharing local data with others, which protect local data privacy.
- Typical pipelines:



Step 3: Aggregate received models to Ws

Step 4: Distribute Ws back to clients for the next round update

Step 2: Upload selected clients' W to the server

Step 1: train local models with local data

W: parameter
M: Model
D: Data

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

Content

1/ Background

2/ Motivation

3/ Proposed Work

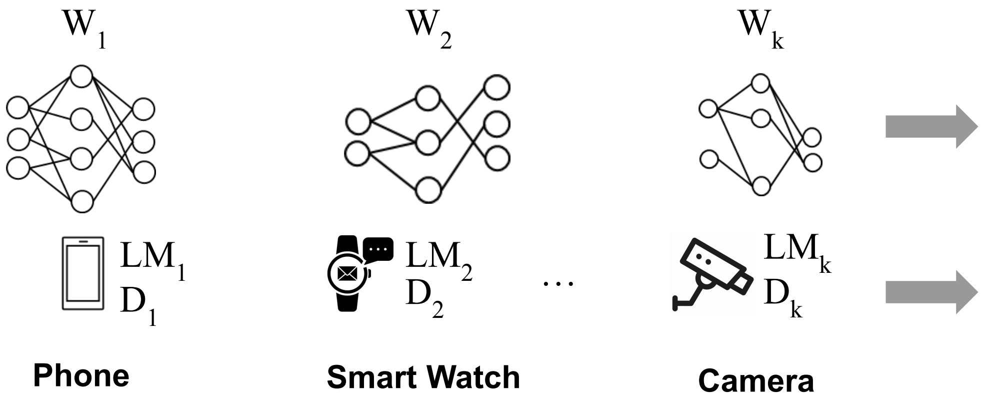
4/ Experiment Results

5/ Conclusion

Motivation and Introduction

- Most FL research works, clients hold fully labeled data. However, labels-at-server scenario is more practical in the real-world setting, where the server holds limited labeled data and clients hold unlabeled data.
- For IoT:
  - Due to the limited computational capability, large model may not be a good fit for local side training.
  - Communication cost is another concern we need to take care of.
- Generalized semi-supervised federated learning framework for image and text data.

W: parameter
LM: Lightweight Model
D: Unlabeled Data



W_1 W_2 W_k

LM_1 D_1 LM_2 D_2 … LM_k D_k

Phone **Smart Watch** **Camera**

Lightweight heterogeneous models

Unlabeled data at clients

How can we design a FL framework to enable lightweight heterogeneous local clients to cooperate effectively under the semi-supervised setting in IoT?

Challenges

- There is no labeled data at the client side, which makes it impossible to directly apply existing personalized FL techniques such as pFedMe[1], since they all need labeled client data.
- To save computational resources and reduce communication cost, IoT applications usually use models with less parameters, i.e., lightweight models. Those models are hard to maintain competitive performance with the large ones. Therefore, the new challenges are how to guarantee the performance of the new SemiFL model and achieve the personalization of client models simultaneously.

[1] T Dinh, Canh, Nguyen Tran, and Josh Nguyen. "Personalized federated learning with moreau envelopes." Advances in Neural Information Processing Systems 33 (2020): 21394-21405.

Related Work

- There are federated learning research works in IoT [1,2], but most of them are conducted with the supervised setting.
- Semi-supervised federated learning works[3,4,5] are also explored, but none of them takes model personalization into consideration. Moreover, only a few models consider communication cost in the model design or the constraints in IoT setting.

[1] Pang, Junjie, et al. "Realizing the heterogeneity: A self-organized federated learning framework for IoT." IEEE Internet of Things Journal 8.5 (2020): 3088-3098.
[2] Li, Zonghang, et al. "Data heterogeneity-robust federated learning via group client selection in industrial IoT." IEEE Internet of Things Journal (2022).
[3] Jeong, Wonyong, et al. "Federated semi-supervised learning with inter-client consistency & disjoint learning." arXiv preprint arXiv:2006.12097 (2020).
[4] Zhang, Zhe, et al. "Semi-supervised federated learning with non-IID data: Algorithm and system design." 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). IEEE, 2021.
[5] Zhang, Zhengming, et al. "Improving semi-supervised federated learning by reducing the gradient diversity of models." 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021.

Content

1 Background

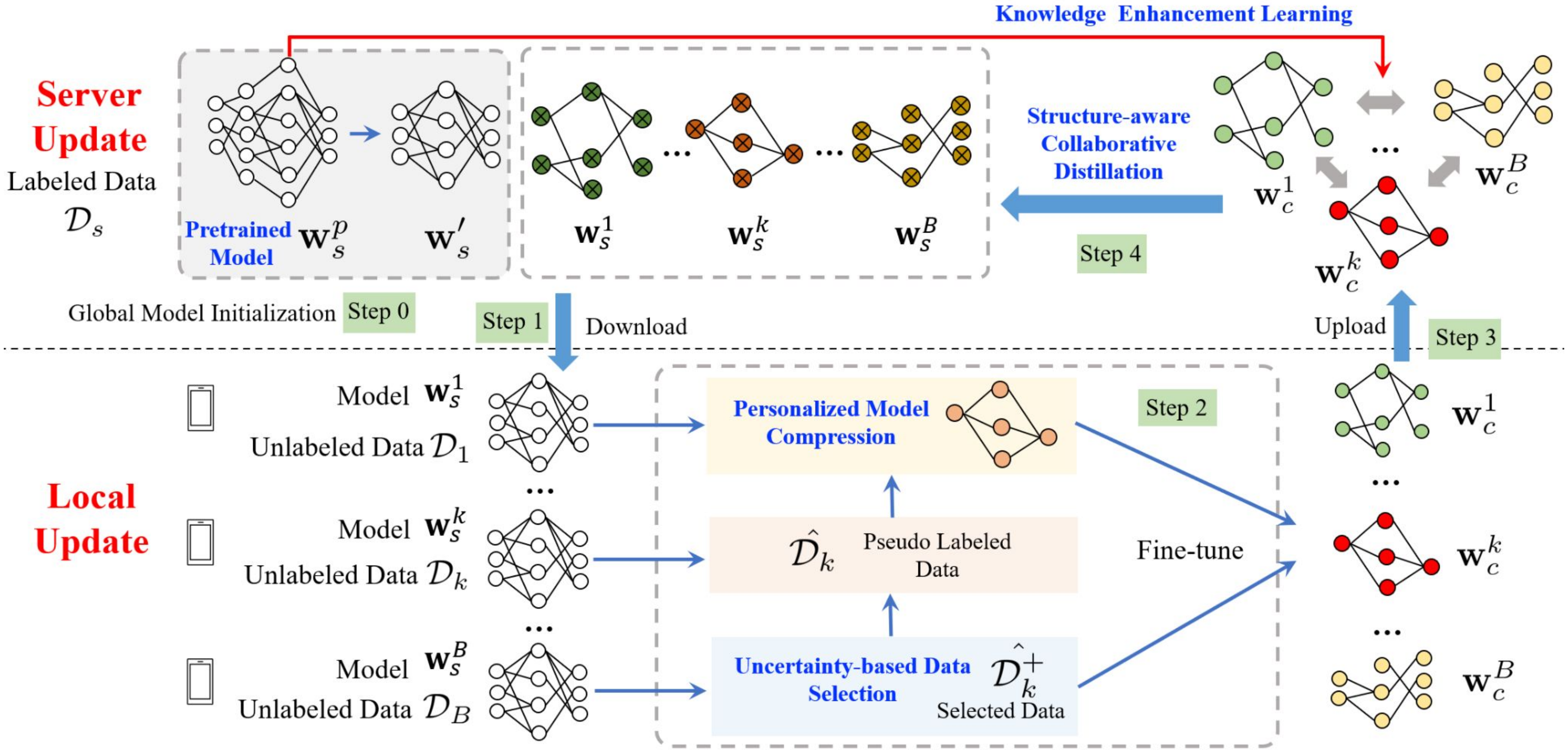
2 Motivation

3 Proposed Work

4 Experiment Results

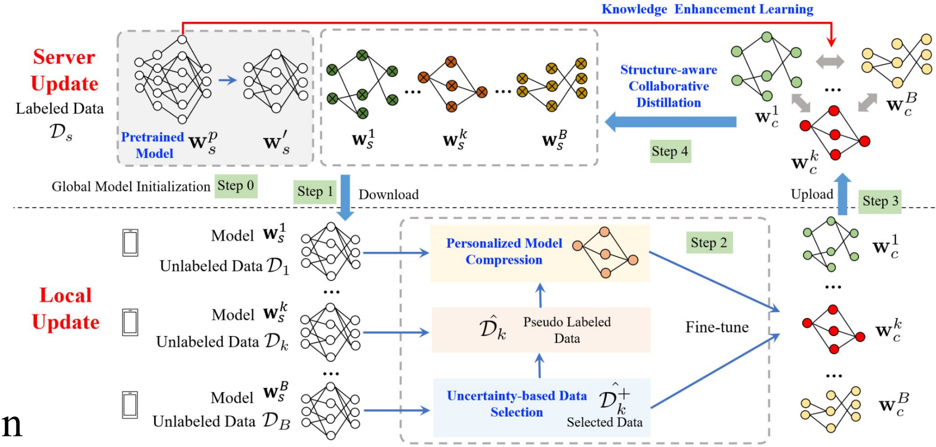
5 Conclusion

Framework Overview

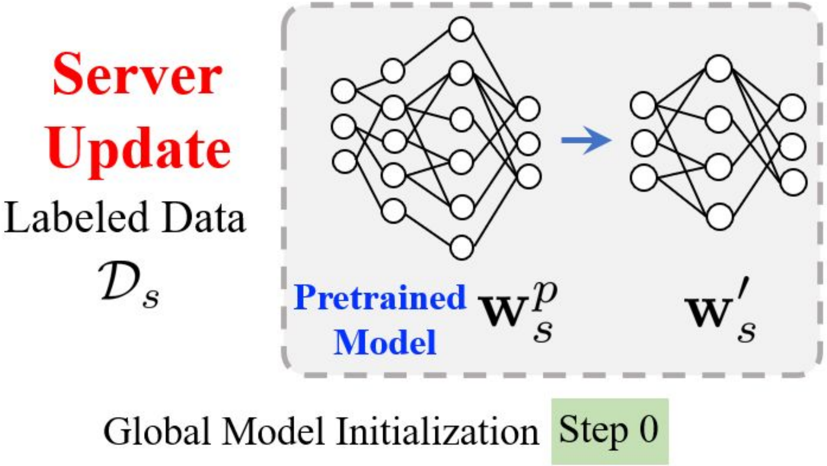


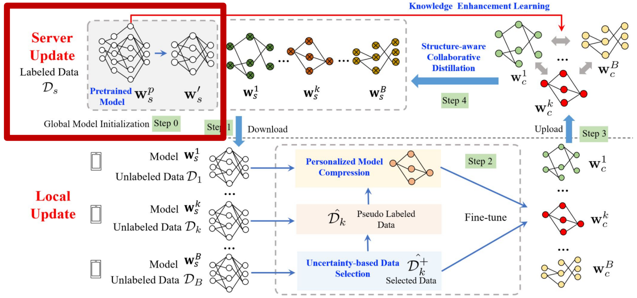
Work Outline

- **System Initialization**
- **Model Personalization (Local Update)**
  - Personalized Model Compression
  - Uncertainty-based Data Selection
  - Personalized Model Update
- **Collaborative Distillation (Server Update)**
  - Structure-aware Similarity Learning
  - Knowledge-enhanced Collaborative Distillation

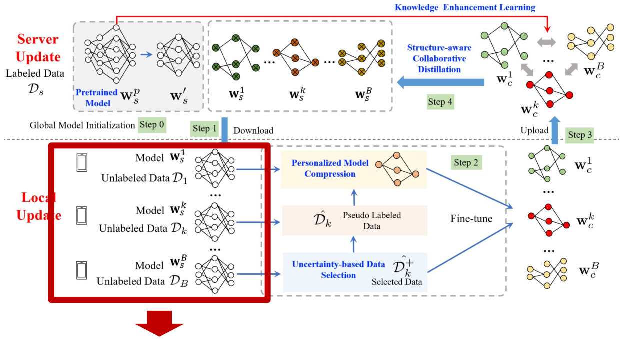


Step 0: System Initialization

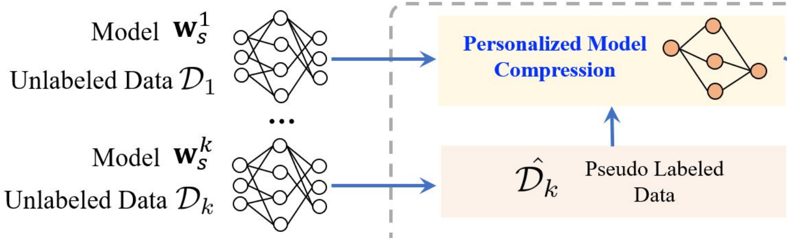




Step 2-1:Personalized Model Compression

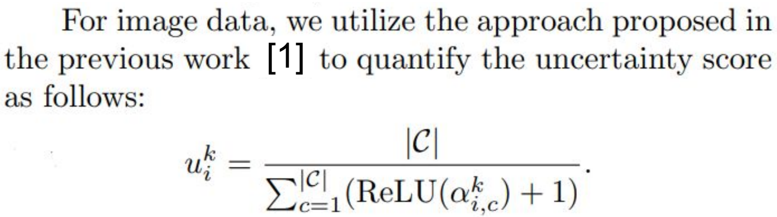

Knowledge Enhancement Learning
Server Update
Labeled Data $\mathcal{D}_s$
Pretrained Model $\mathbf{w}_s^p$
$\mathbf{w}'_s$
$\mathbf{w}_s^1$
$\mathbf{w}_s^k$
$\mathbf{w}_s^B$
Structure-aware Collaborative Distillation
Step 4
$\mathbf{w}_c^1$
$\mathbf{w}_c^k$
$\mathbf{w}_c^B$
Global Model Initialization
Step 0
Step 1
Download
Upload
Step 3
Local Update
Model $\mathbf{w}_s^1$
Unlabeled Data $\mathcal{D}_1$
Model $\mathbf{w}_s^k$
Unlabeled Data $\mathcal{D}_k$
Model $\mathbf{w}_s^B$
Unlabeled Data $\mathcal{D}_B$
Personalized Model Compression
Step 2
$\hat{\mathcal{D}}_k$ Pseudo Labeled Data
Fine-tune
Uncertainty-based Data Selection
$\hat{\mathcal{D}}_k^+$ Selected Data
$\mathbf{w}_c^1$
$\mathbf{w}_c^k$
$\mathbf{w}_c^B$

- Generate pseudo labels with the received global model
- Model compression to generate customized local models


Model $\mathbf{w}_s^1$
Unlabeled Data $\mathcal{D}_1$
Model $\mathbf{w}_s^k$
Unlabeled Data $\mathcal{D}_k$
Personalized Model Compression
$\hat{\mathcal{D}}_k$
Pseudo Labeled Data

Step 2-2: Uncertainty-based Data Selection

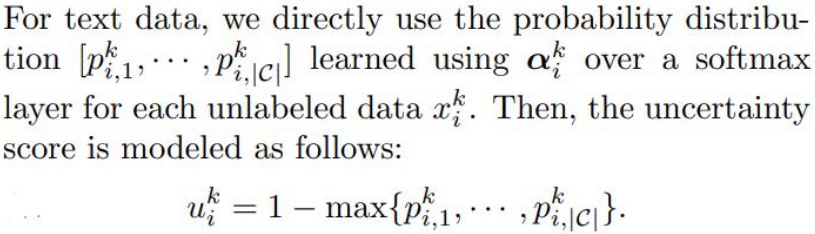
- We select high-quality pseudo labels to fine-tune the compressed models
- For image data:



For image data, we utilize the approach proposed in the previous work [1] to quantify the uncertainty score as follows:

$$u_i^k = \frac{|\mathcal{C}|}{\sum_{c=1}^{|\mathcal{C}|}(\mathrm{ReLU}(\alpha_{i,c}^k)+1)}.$$

- For text data:



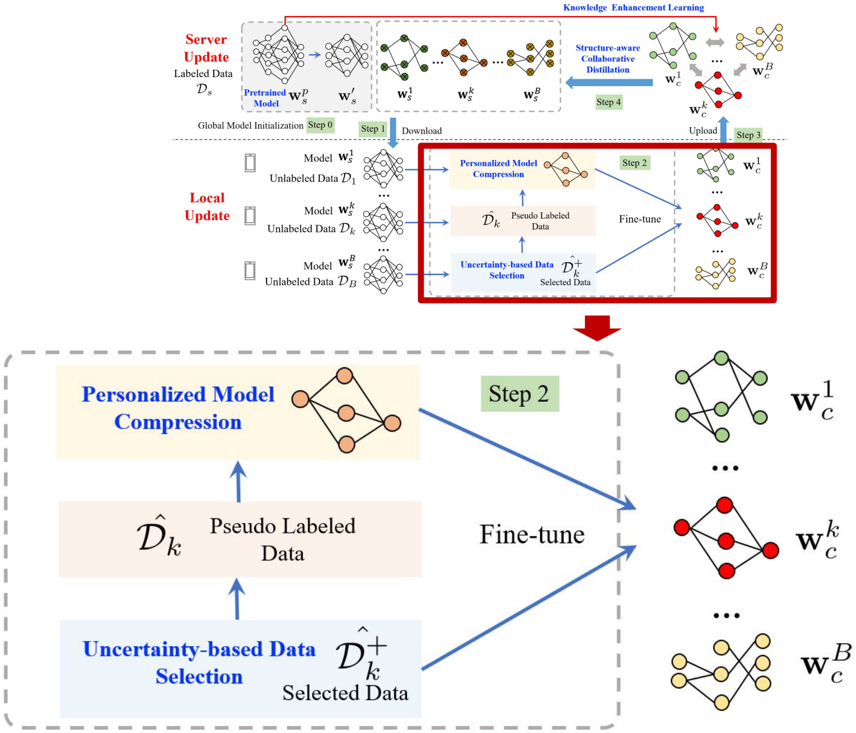
For text data, we directly use the probability distribution $[p_{i,1}^k, \cdots, p_{i,|\mathcal{C}|}^k]$ learned using $\boldsymbol{\alpha}_i^k$ over a softmax layer for each unlabeled data x_i^k. Then, the uncertainty score is modeled as follows:

$$u_i^k = 1 - \max\{p_{i,1}^k, \cdots, p_{i,|\mathcal{C}|}^k\}.$$

[1] Sensoy, Murat, Lance Kaplan, and Melih Kandemir. "Evidential deep learning to quantify classification uncertainty." Advances in neural information processing systems 31 (2018).

Step 2-3:Personalized Model Update

- Using selected high-quality pseudo labeled data, we fine tune the lightweight mode.
- Upload to the server (step 3).

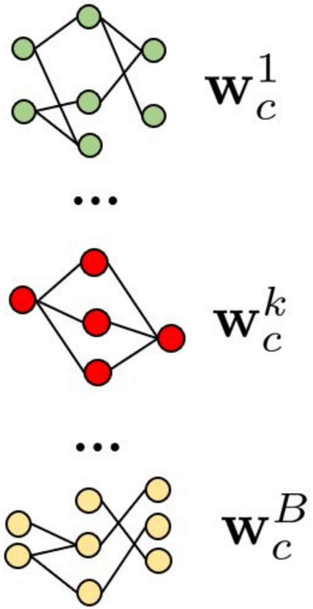


Till now, we have finished the initialization and local updates.
We obtain personalized local models with customized parameters and structures.

However, we face new challenges here.

Raised Challenges at the Server Side

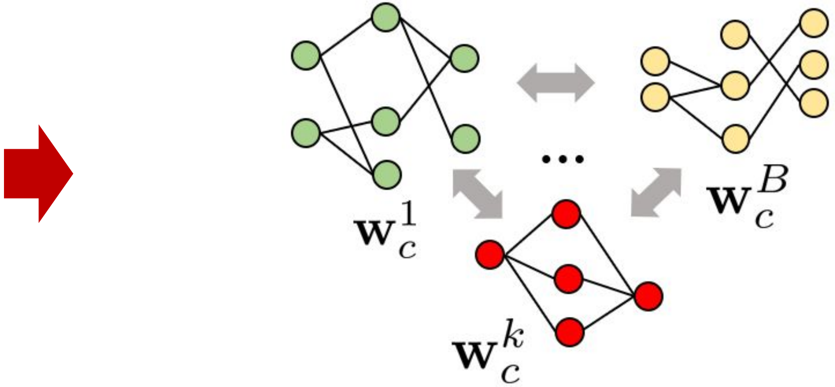
- For personalized learning, we need to maintain the key characteristics of the current model as well as taking other uploaded model parameters into consideration to absorb common knowledge for further enhancing the learning.
- Federated learning needs to train each local model iteratively. Thus, it is difficult for the training of models to converge if the model architectures change frequently and dramatically.
- As each local model is a sub-structure of the original model, which may cause compact model to miss the general knowledge that is learned by the original pretrained model.



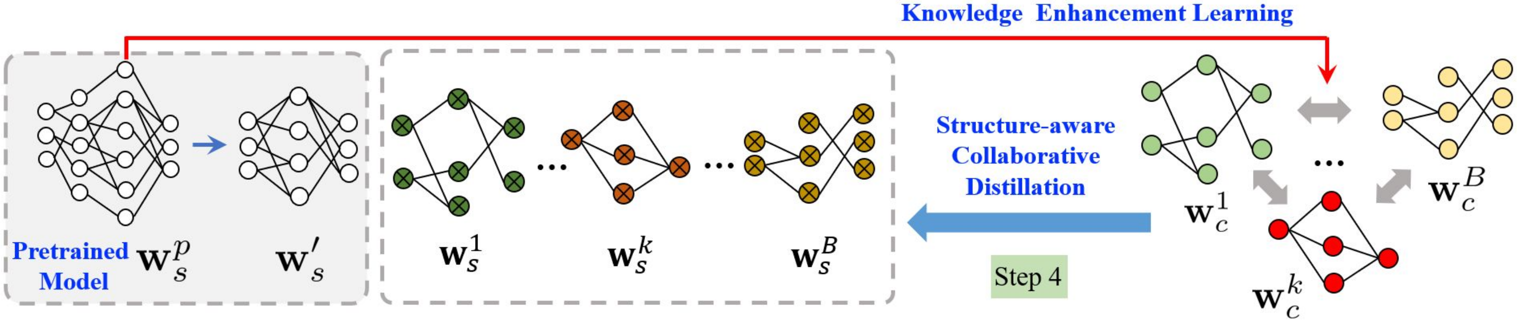
Thus, it requires that each model must keep its original architecture during the model fusion as well as aggregating appropriate knowledge from other clients and general knowledge from the large model stored at the server side.

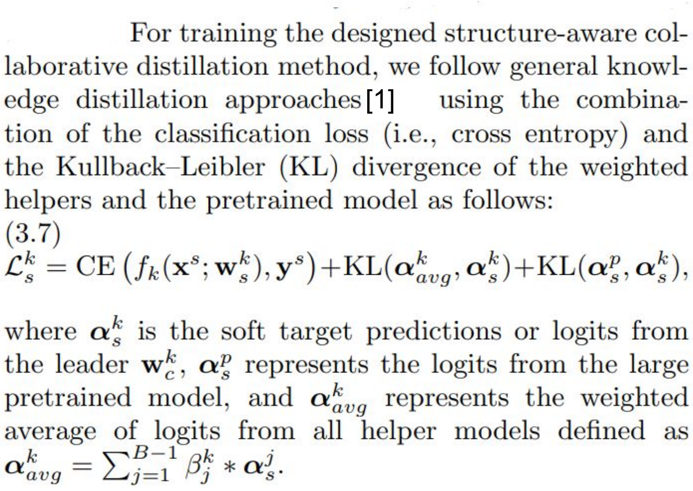
Structure-aware Similarity Learning

- Intuitively, if the data distributions on two different clients are similar, the lightweight client models will also have similar network structures, which further generate similar outputs.
- Thus, different helpers will contribute differently, and it is essential to distinguish the importance of helpers.



Step 4: Knowledge-Enhanced Collaborative Distillation



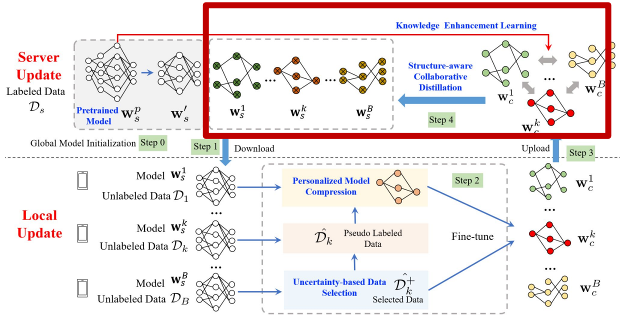


For training the designed structure-aware collaborative distillation method, we follow general knowledge distillation approaches [1] using the combination of the classification loss (i.e., cross entropy) and the Kullback–Leibler (KL) divergence of the weighted helpers and the pretrained model as follows:

(3.7)

$$\mathcal{L}_s^k = \text{CE}\left(f_k(\mathbf{x}^s;\mathbf{w}_s^k),\mathbf{y}^s\right)+\text{KL}(\boldsymbol{\alpha}_{avg}^k,\boldsymbol{\alpha}_s^k)+\text{KL}(\boldsymbol{\alpha}_s^p,\boldsymbol{\alpha}_s^k),$$

where $\boldsymbol{\alpha}_s^k$ is the soft target predictions or logits from the leader $\mathbf{w}_c^k$, $\boldsymbol{\alpha}_s^p$ represents the logits from the large pretrained model, and $\boldsymbol{\alpha}_{avg}^k$ represents the weighted average of logits from all helper models defined as $\boldsymbol{\alpha}_{avg}^k = \sum_{j=1}^{B-1}\beta_j^k * \boldsymbol{\alpha}_s^j$.



[1] Zhang, Ying, et al. "Deep mutual learning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Content

1/ Background

2/ Motivation

3/ Proposed Work

4/ Experiment Results

5/ Conclusion

Performance Evaluation and Communication Efficiency Analysis

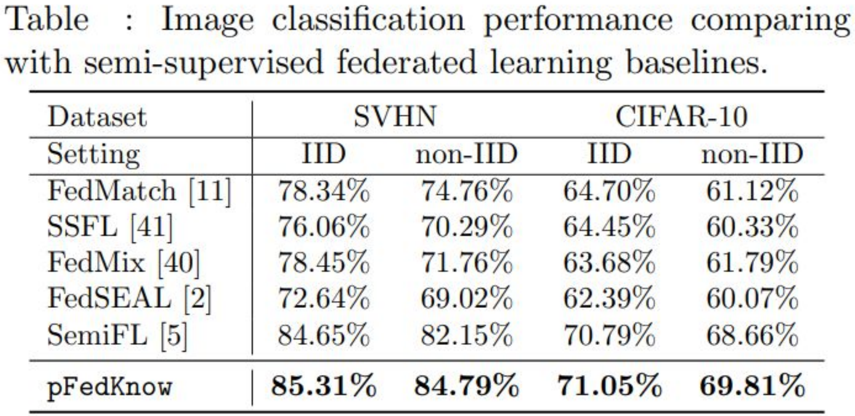


Table : Image classification performance comparing with semi-supervised federated learning baselines.

Dataset	SVHN		CIFAR-10	
Setting	IID	non-IID	IID	non-IID
FedMatch [11]	78.34%	74.76%	64.70%	61.12%
SSFL [41]	76.06%	70.29%	64.45%	60.33%
FedMix [40]	78.45%	71.76%	63.68%	61.79%
FedSEAL [2]	72.64%	69.02%	62.39%	60.07%
SemiFL [5]	84.65%	82.15%	70.79%	68.66%
pFedKnow	**85.31%**	**84.79%**	**71.05%**	**69.81%**

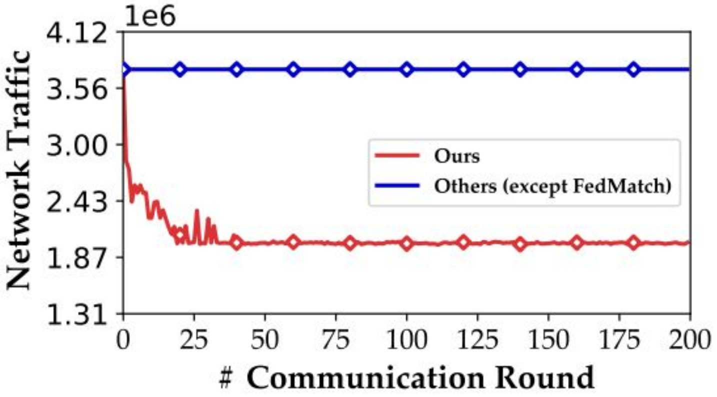
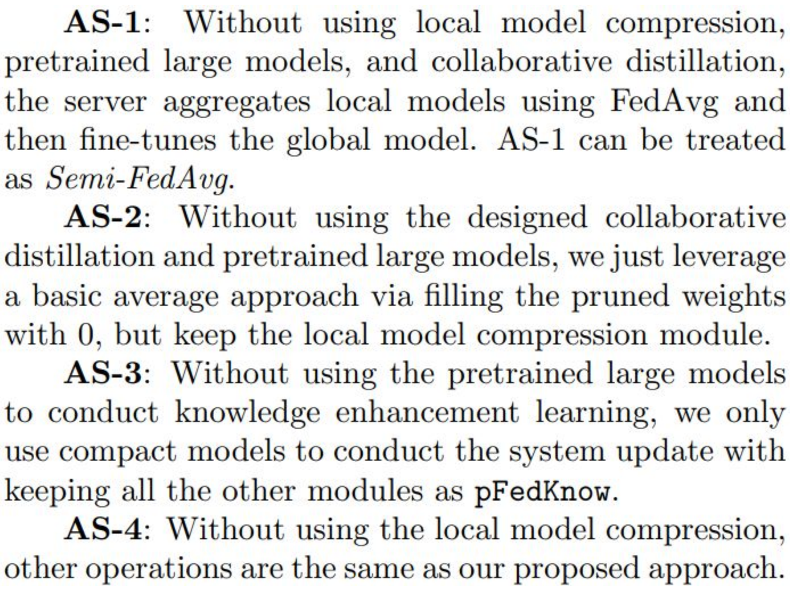


Figure : Communication cost analysis on CIFAR-10.

Ablation Study



AS-1: Without using local model compression, pretrained large models, and collaborative distillation, the server aggregates local models using FedAvg and then fine-tunes the global model. AS-1 can be treated as *Semi-FedAvg*.

AS-2: Without using the designed collaborative distillation and pretrained large models, we just leverage a basic average approach via filling the pruned weights with 0, but keep the local model compression module.

AS-3: Without using the pretrained large models to conduct knowledge enhancement learning, we only use compact models to conduct the system update with keeping all the other modules as `pFedKnow`.

AS-4: Without using the local model compression, other operations are the same as our proposed approach.

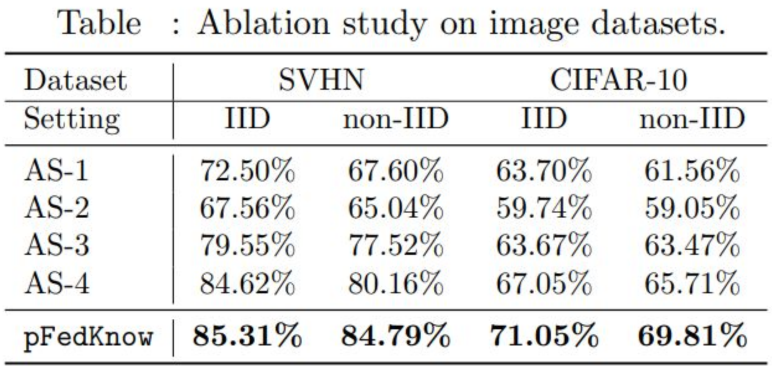
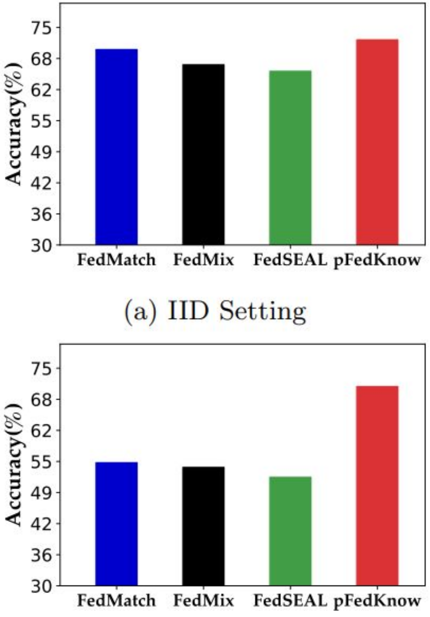


Table : Ablation study on image datasets.

Dataset	SVHN		CIFAR-10	
Setting	IID	non-IID	IID	non-IID
AS-1	72.50%	67.60%	63.70%	61.56%
AS-2	67.56%	65.04%	59.74%	59.05%
AS-3	79.55%	77.52%	63.67%	63.47%
AS-4	84.62%	80.16%	67.05%	65.71%
`pFedKnow`	**85.31%**	**84.79%**	**71.05%**	**69.81%**

Selected Experimental Results on Text Data



(a) IID Setting

(b) Non-IID Setting

Figure : Performance comparison on AG.

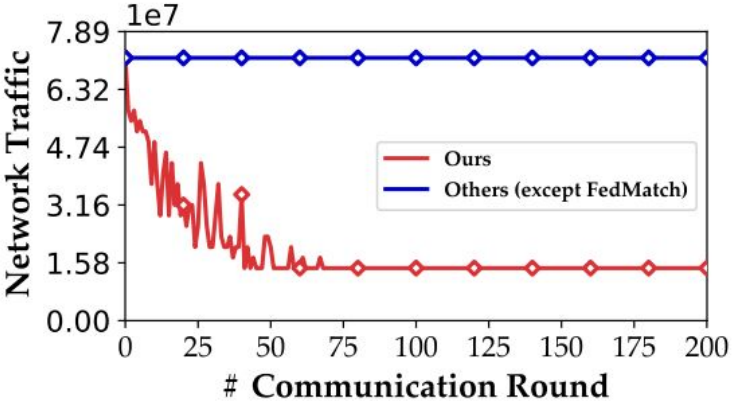
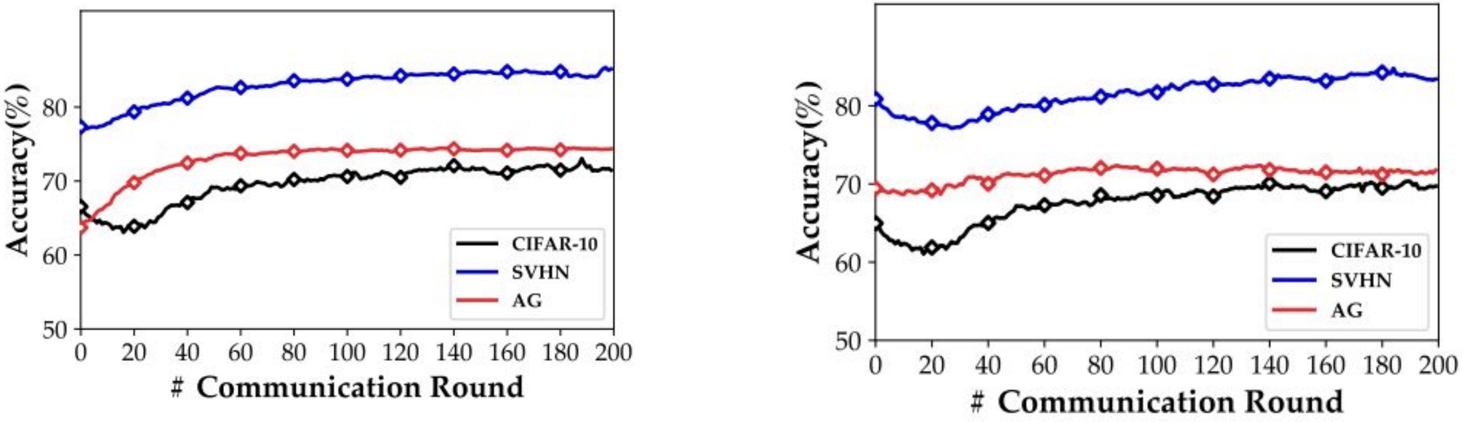


Figure : Communication cost analysis on AG.

Image and Text Convergence



(a) IID Setting.

(b) Non-IID Setting.

Figure : Image and text convergence analysis.

Content

1/ Background

2/ Motivation

3/ Proposed Work

4/ Experiment Results

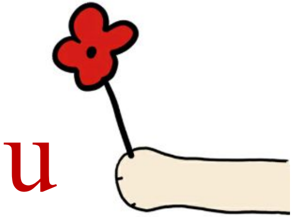
5/ Conclusion

Contribution Summary

- To the best of our knowledge, we are the first work to distill lightweight models to warm up and further customize compressed local models with different structures using network pruning techniques in FL, which further solves the challenges of the limited local device computational capacity and restricted network bands in IoT.
- We propose a new aggregation approach with the combination of network structure-aware collaborative distillation and large-model knowledge enhancement learning, which can obtain a personalized model with the help of other models even with different structures and extract general knowledge from pretrained large models.
- We conduct extensive experiments on both image and text datasets to show the effectiveness and efficiency of the proposed framework compared with state-of-the-art baselines.

Any Questions or Comments?

Thank you



Jiaqi Wang: jqwang@psu.edu
Fenglong Ma: fenglong@psu.edu
PSU Data Science Lab
https://psudslab.github.io/