Personalized Federated Learning over Heterogeneous Data

– Jun Luo's PhD Proposal Defense
November 3rd , 2025
Intelligent Systems Program
University of Pittsburgh

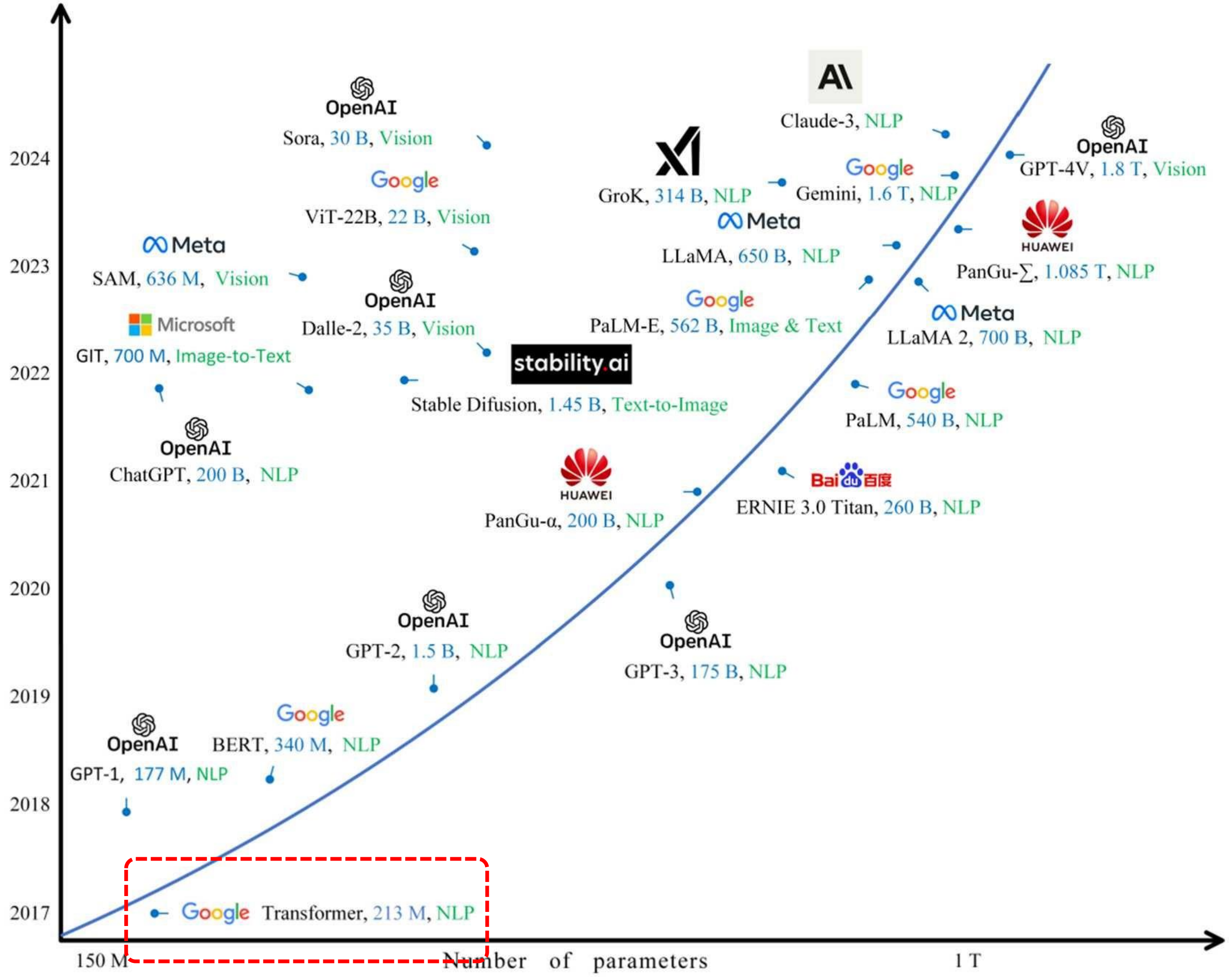
Committee

- Chair
  - Dr. Shandong Wu, Associate Professor, Intelligent Systems Program
- Members
  - Dr. Leming Zhou, Associate Professor, Intelligent Systems Program
  - Dr. Xiaowei Jia, Assistant Professor, Intelligent Systems Program
  - Dr. Lu Tang, Assistant Professor, Department of Biostatistics

List of Publications

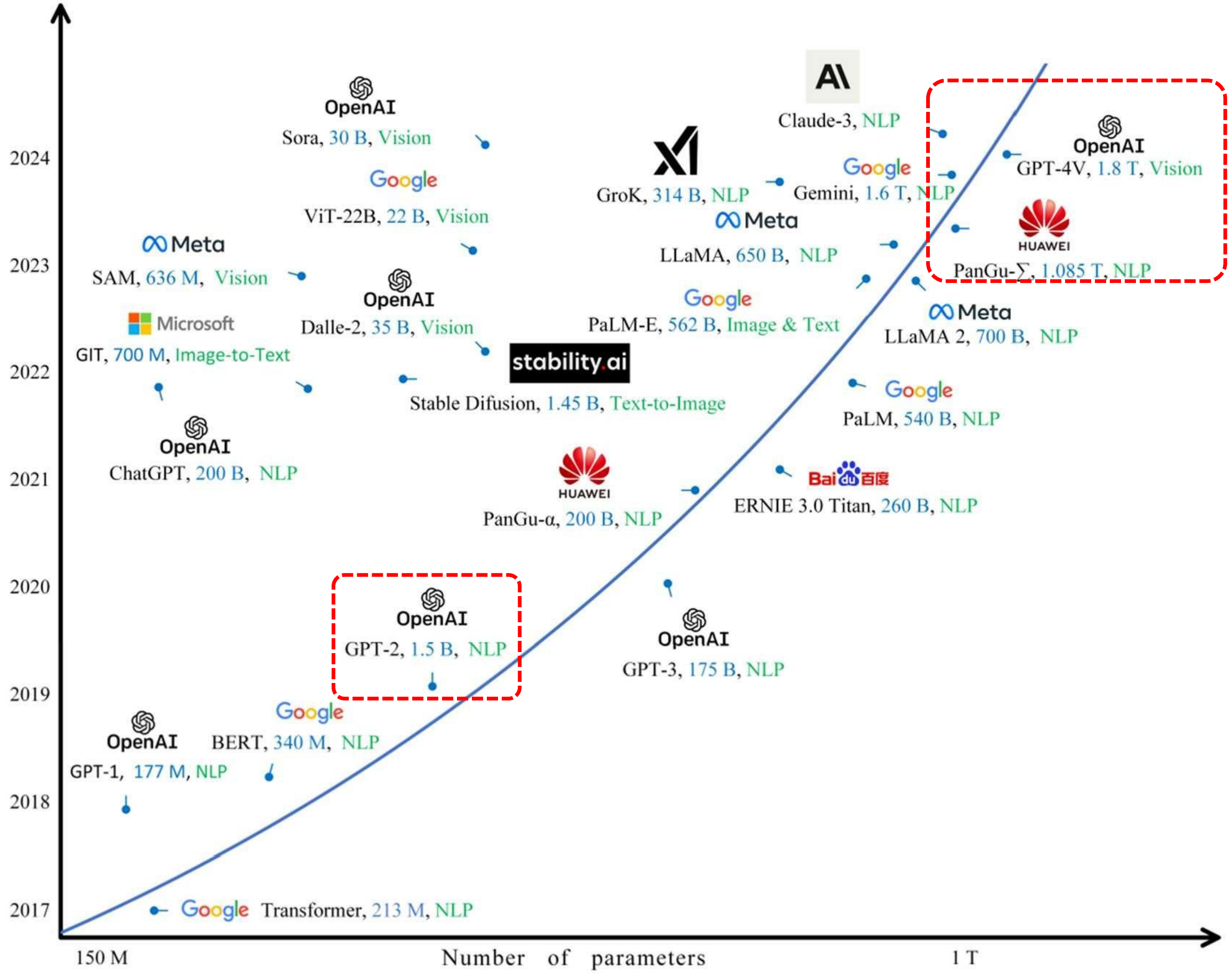
1. **Jun Luo**, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *Proceedings of the Thirteenth International Conference on Learning Representation*, 2025
2. **Jun Luo**, Matias Mendieta, Chen Chen, and Shandong Wu. Pgfed: Personalize each client's global objective for federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3946–3956, 2023.
3. **Jun Luo** and Shandong Wu. Adapt to adaptation: Learning personalization for cross-silo federated learning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2166–2173. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.
4. **Jun Luo** and Shandong Wu. Fedsld: Federated learning with shared label distribution for medical image classification. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.
5. **Jun Luo**, Dooman Arefan, Margarita Zuley, Jules Sumkin, and Shandong Wu. Deep curriculum learning in task space for multi-class based mammography diagnosis. In *Medical Imaging 2022: Computer-Aided Diagnosis*, volume 12033, pages 85–90. SPIE, 2022.
6. **Jun Luo**, Gene Kitamura, Dooman Arefan, Emine Doganay, Ashok Panigrahy, and Shandong Wu. Knowledge-guided multiview deep curriculum learning for elbow fracture classification. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021*, Strasbourg, France, September 27, 2021, Proceedings 12, pages 555–564. Springer, 2021.
7. **Jun Luo**, Gene Kitamura, Emine Doganay, Dooman Arefan, and Shandong Wu. Medical knowledge-guided deep curriculum learning for elbow fracture diagnosis from x-ray images. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 247–252. SPIE, 2021.
8. **Jun Luo**, Dooman Arefan, Margarita Zuley, Chen Chen, Jules Sumkin, Shandong Wu. Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection. *In the Radiological Society of North America (RSNA) Annual Meeting (abstract)*, 2025
9. **Jun Luo**, Dooman Arefan, Anil Vasireddi, Shandong Wu, and Nghi Nguyen. Potential use of artificial intelligence in sincalide-stimulated cholescintigraphy: A pilot study. In *Society of Nuclear Medicine and Molecular Imaging (SNMMI) Annual Meeting* (abstract, full paper under review by European Journal of Nuclear Medicine and Molecular Imaging), 2023.
10. Guangyu Sun, Matias Mendieta, **Jun Luo**, ShandongWu, and Chen Chen. Fedperfix: Towards partial model personalization of vision transformers in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4988–4998, 2023.
11. Zhengbo Zhou, **Jun Luo**, Dooman Arefan, Gene Kitamura, and Shandong Wu. Human not in the loop: objective sample difficulty measures for curriculum learning. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
12. Suvendra Vijayan, **Jun Luo**, Shandong Wu, and Anitha Potluri. Image enhancement of ultralow dose cbct images using a deep generative model. In *American Roentgen Ray Society (ARRS) Annual Meeting*, 134(3):e72 (abstract), 2022.
13. Emine Doganay, Gene Kitamura, Luo Yang, **Jun Luo**, and ShandongWu. Multi-view-enabled deep learning for automated radiographic view classification and fracture detection of the elbow. In *American Roentgen Ray Society (ARRS) Annual Meeting* (abstract). 2021.

Background



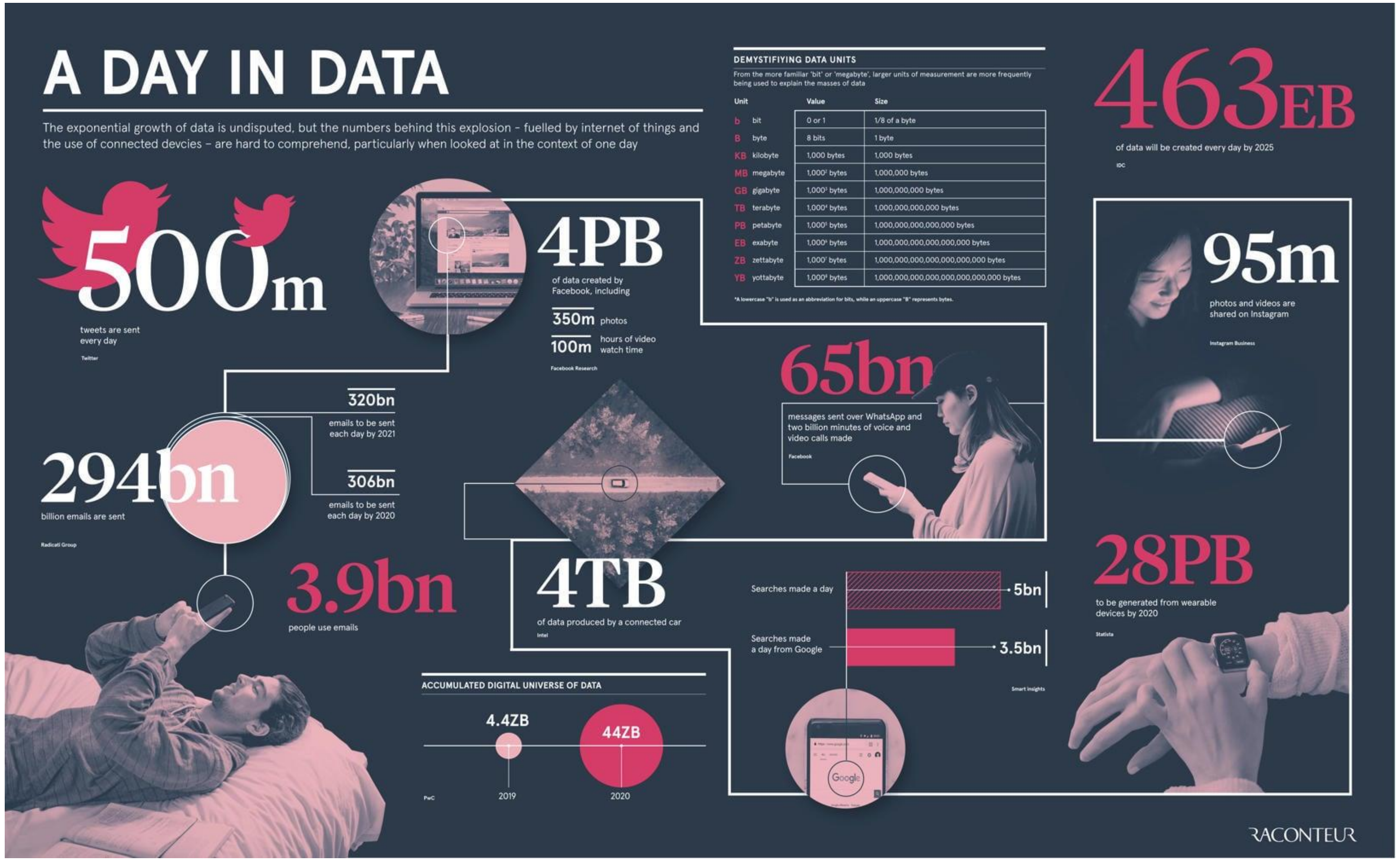
[Tu et. al., 2024]

Background

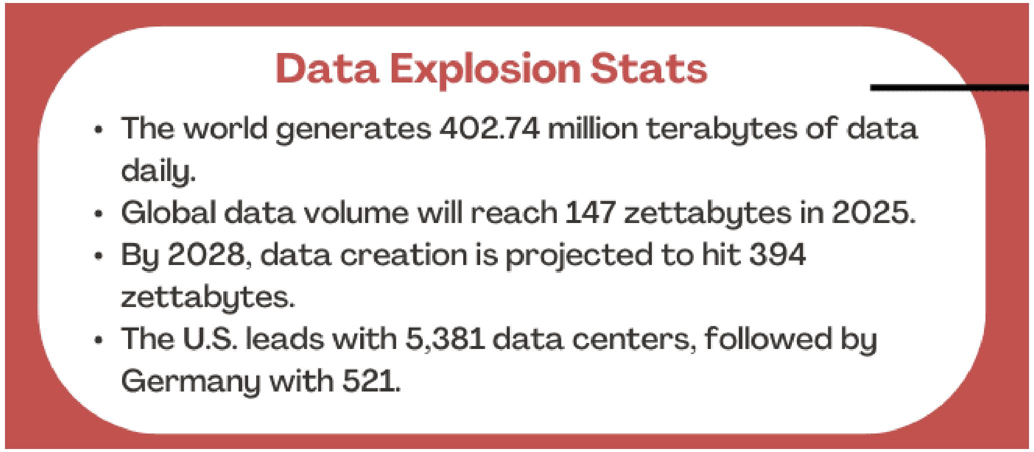


[Tu et. al., 2024]

Background



[Jeff Desjardins, World Economic Forum]



Data Explosion Stats

- The world generates 402.74 million terabytes of data daily.
- Global data volume will reach 147 zettabytes in 2025.
- By 2028, data creation is projected to hit 394 zettabytes.
- The U.S. leads with 5,381 data centers, followed by Germany with 521.

[Khyati Hooda, KeyWordsEverywhere.com]

Background

Internet of things data

[Risk Management Magazine]

Health data

[American Hospital Association]

Background

SPRINTO Frameworks Platform Resources

Blog > Blogs > GDPR Privacy Policy: Ensuring Compliance with EU Data Rules

GDPR Privacy Policy: Ensuring Compliance with EU Data Rules

Bhuvesh Lal • Sep 30, 2024

AMA Join Renew Search

HIPAA

HIPAA privacy rule

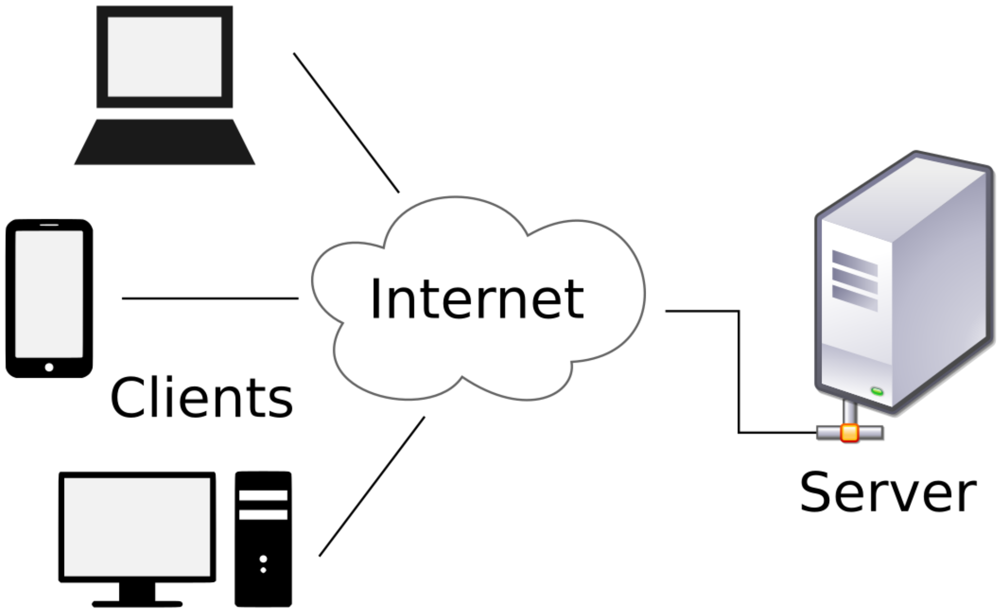
2 Min Read

Save Copy Print Share

The HIPAA Privacy Rule provides federal standards to safeguard the privacy of personal health information and gives patients an array of rights with respect to that information, including rights to examine and obtain a copy of their health records and to request corrections. The U.S. Department of Health & Human Services' (HHS) Office of Civil Rights (OCR) oversees compliance with HIPAA privacy requirements.

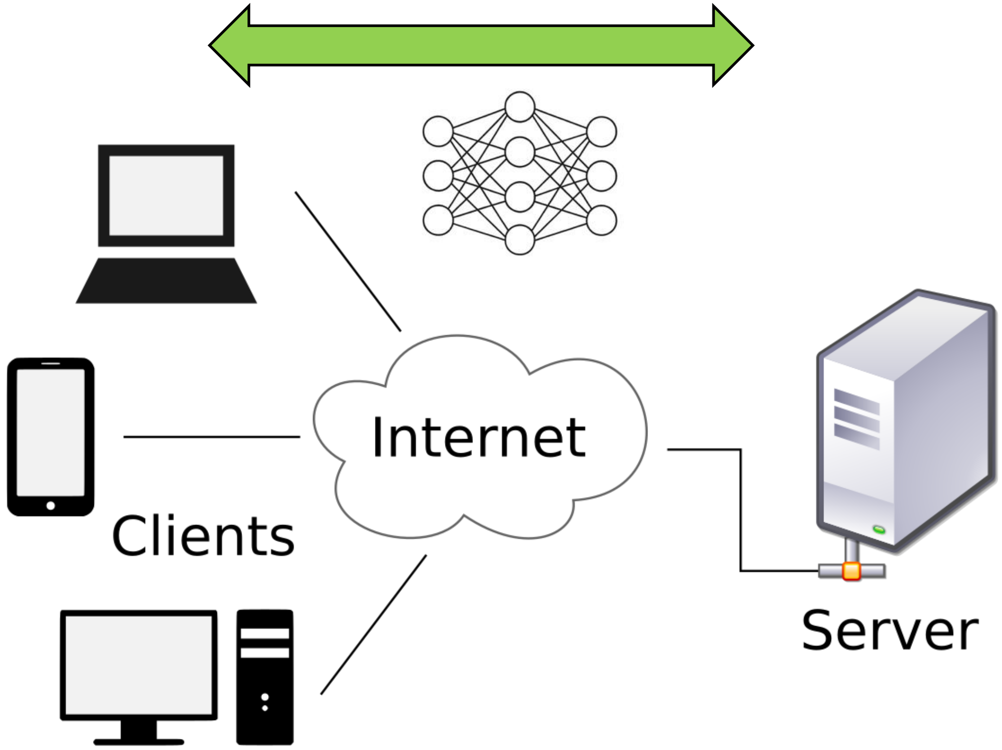
Background

- Federated learning



Background

- Federated learning



Overview

- Federated learning: introduction

- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

- Summary

Overview

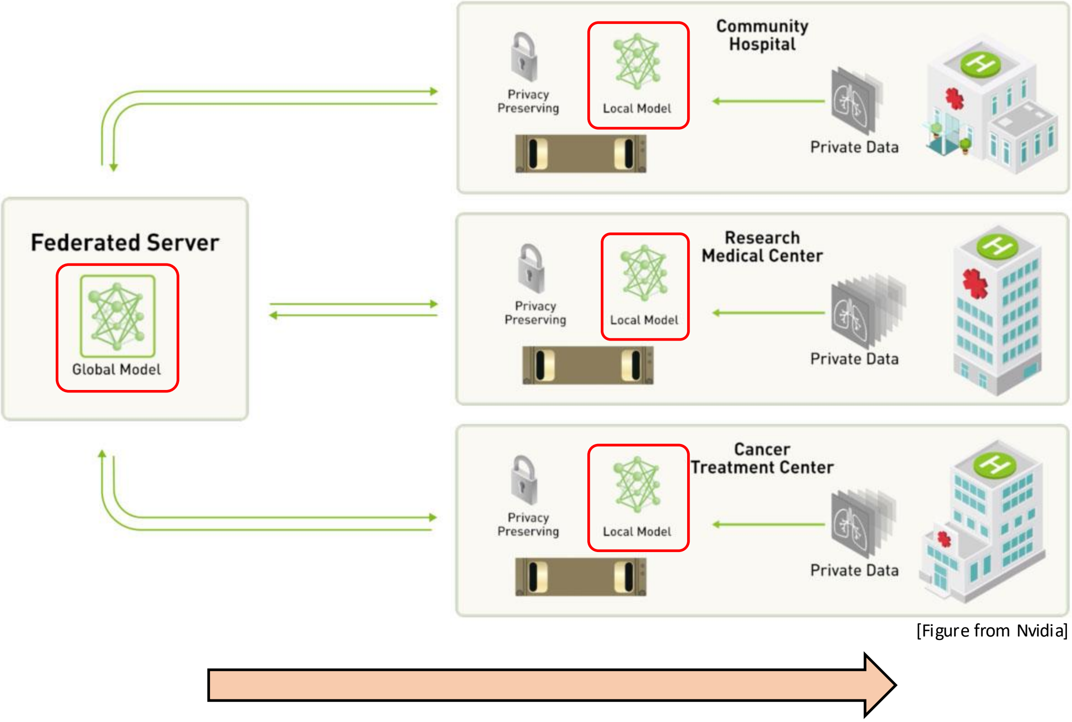
- Federated learning: introduction

- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

- Summary

Federated learning: introduction

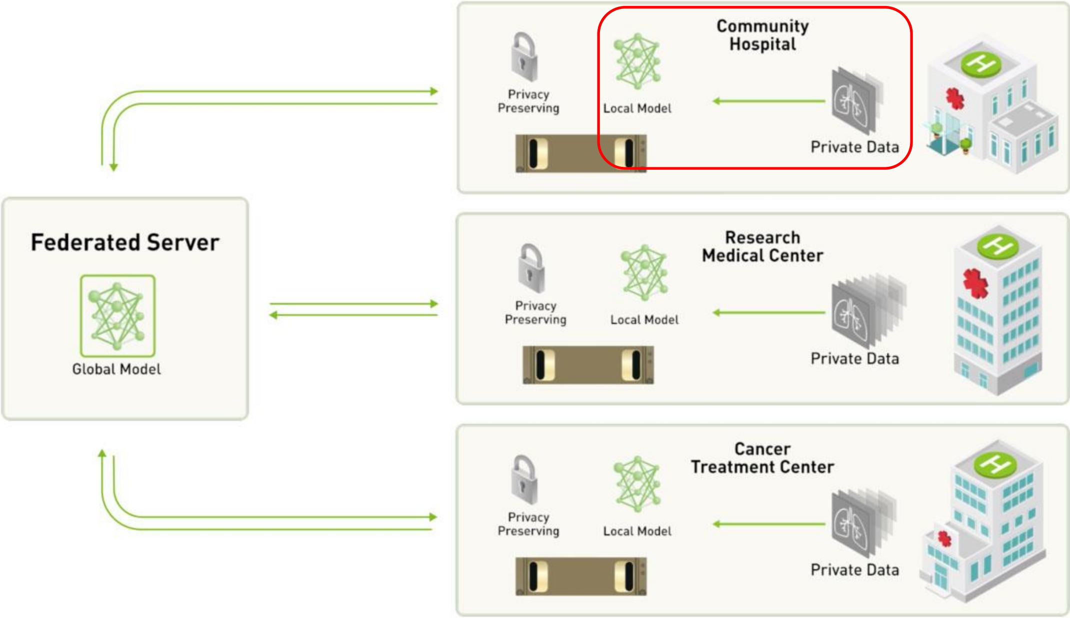
- Basic mechanism of traditional FL
  - Broadcast



[Figure from Nvidia]

Federated learning: introduction

- Basic mechanism of traditional FL
  - Broadcast
  - Client (local) training

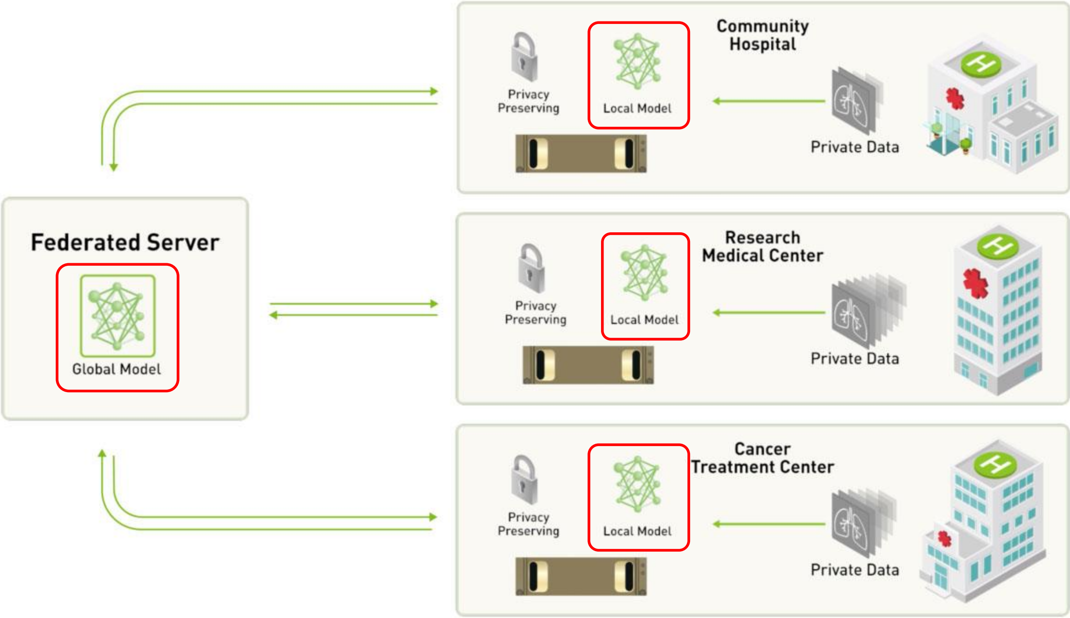


[Figure from Nvidia]

Federated learning: introduction

- Basic mechanism of traditional FL
    - Broadcast
    - Client (local) training
    - Server (global) aggregation

      FedAvg: $w = \sum_i p_i w_i$

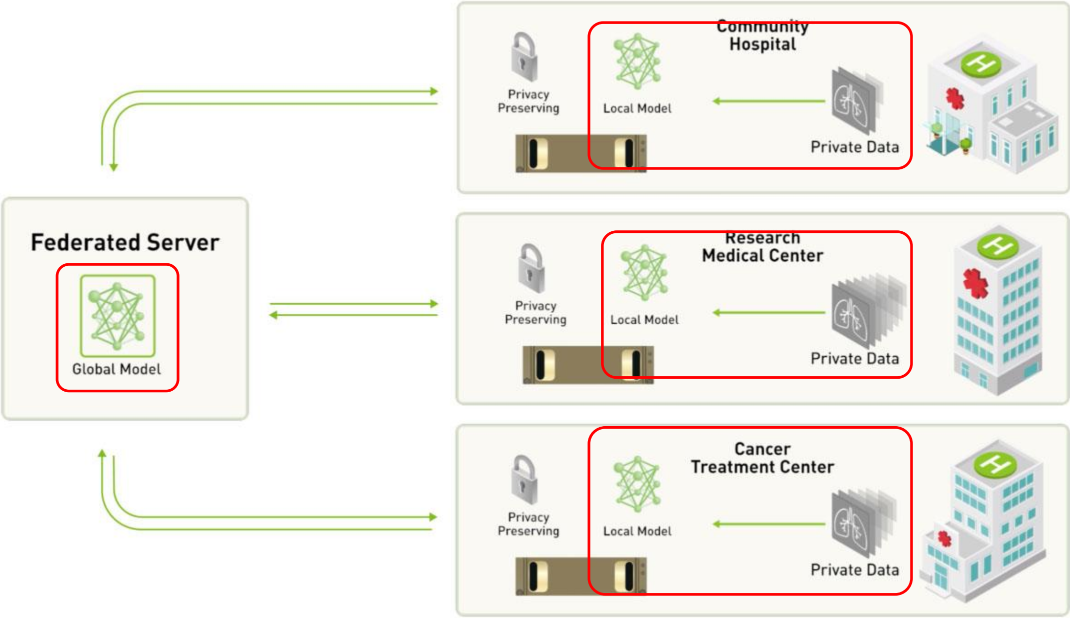


[Figure from Nvidia]

Federated learning: introduction

- Basic mechanism of traditional FL
  - Broadcast
  - Client (local) training
  - Server (global) aggregation

    FedAvg: $w = \sum_i p_i w_i$



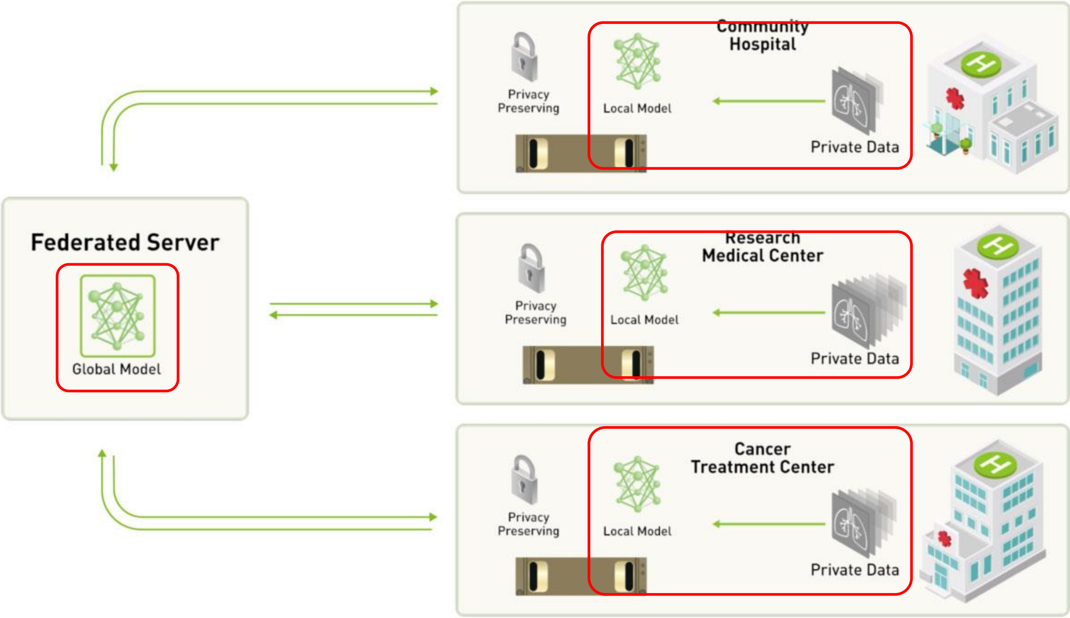
[Figure from Nvidia]

Federated learning: introduction

- Basic mechanism of traditional FL
  - Broadcast
  - Client (local) training
  - Server (global) aggregation

    FedAvg: $w = \sum_i p_i w_i$

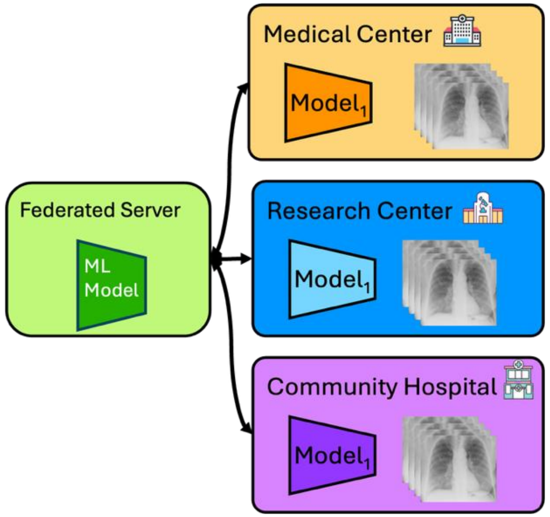
If #clients is large, sample clients at the beginning of each round



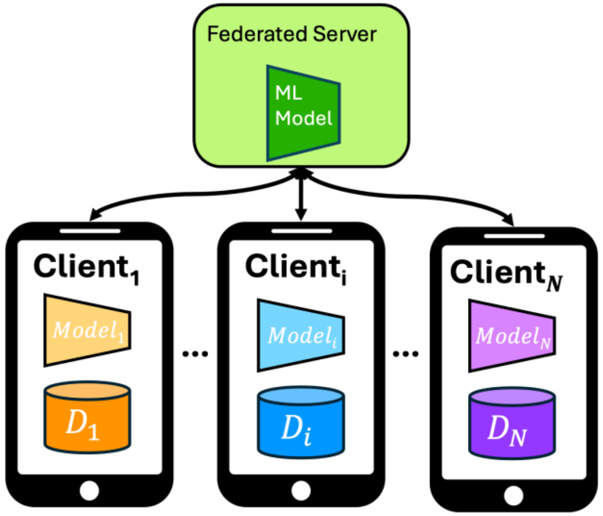
[Figure from Nvidia]

Federated learning: introduction

- Applications of FL
  - Cross-silo FL
    - Medical centers
    - Financial institutes
  - Cross-device FL
    - Smart phone/IoT devices
    - Smart vehicle



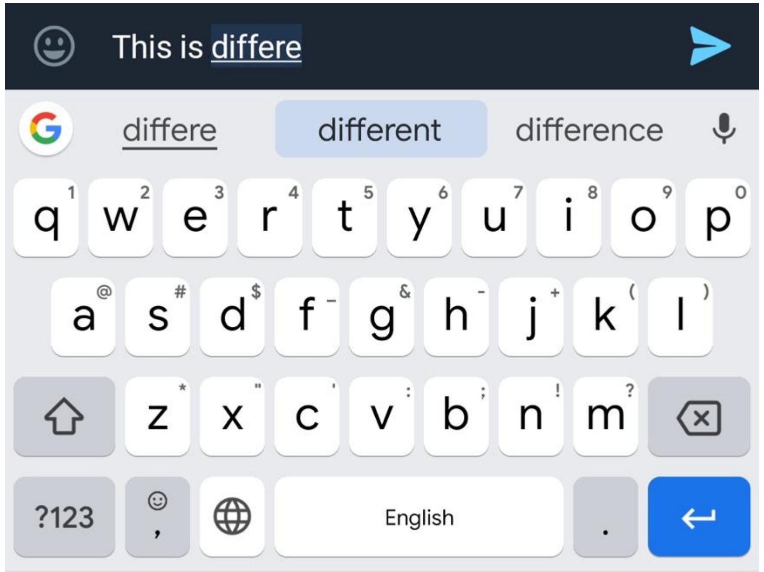
Cross-silo FL



Cross-device FL

Federated learning: introduction

- Applications of FL
  - Cross-silo FL
    - Medical centers
    - Financial institutes
  - Cross-device FL
    - Smart phone/IoT devices
    - Smart vehicle
  - One of the earliest successes of FL: Gboard
  - “Hey Siri” from Apple, “Alexa” from Amazon…



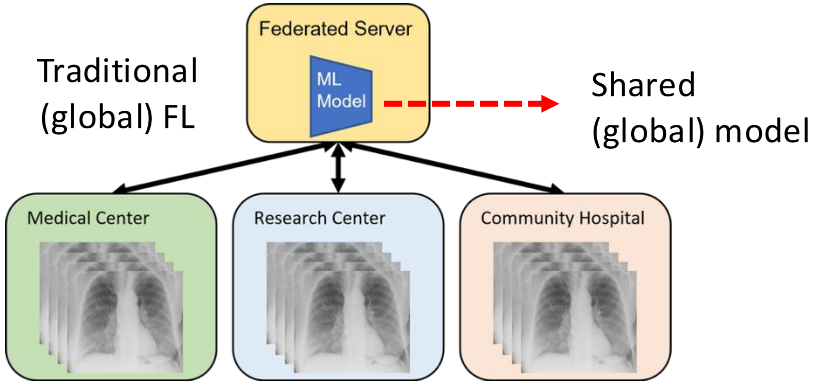
[Figure from Google]

“Hey Siri”

“Alexa”

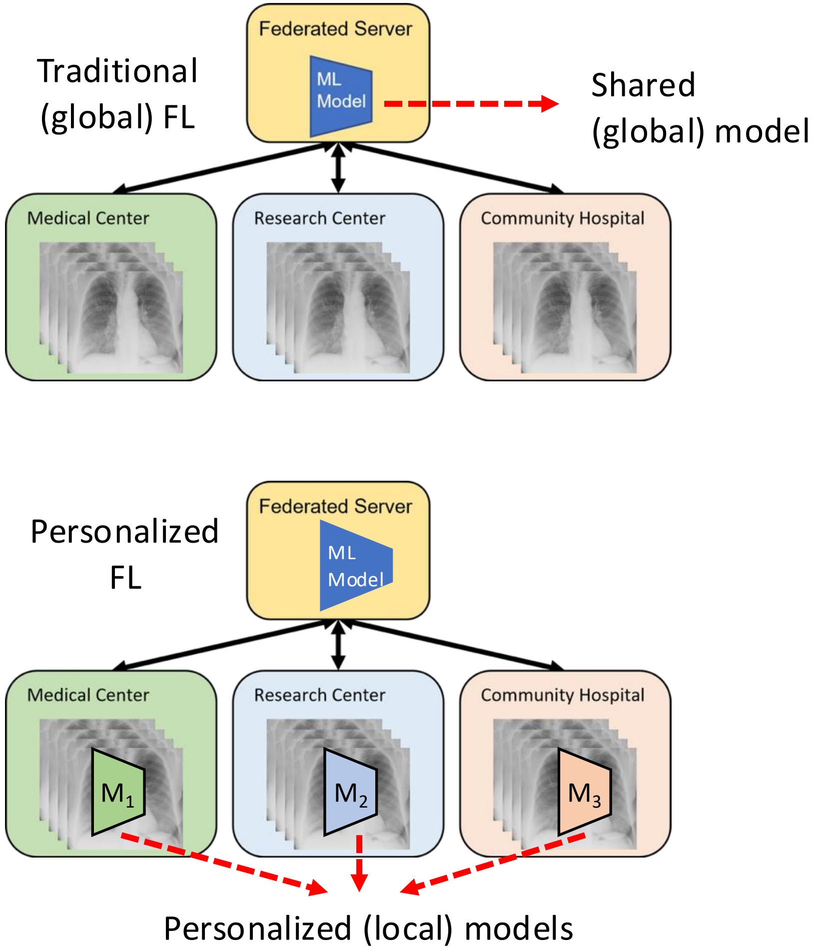
Federated learning: introduction

- Data heterogeneity and personalized FL (PFL)
  - Data heterogeneity – non-IID
    - E.g. medical datasets are often non-IID
      - Different data acquisition protocols
      - Different local demographics



Federated learning: introduction

- Data heterogeneity and personalized FL (PFL)
  - Data heterogeneity – non-IID
    - E.g. medical datasets are often non-IID
      - Different data acquisition protocols
      - Different local demographics
  - Traditional (global) FL
    - Trains a single global consensus model
    - Issues caused by data heterogeneity
      - inferior performance
      - slower convergence
      - Loss of clients' incentives to participate in FL
  - Personalized FL (PFL)
    - Allows customized models for different clients
    - Systemically mitigates data heterogeneity issue



Overview

- Federated learning: introduction

Global FL
- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)

PFL
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)

- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

- Summary

Overview

- Federated learning: introduction

Global FL
ISBI '22

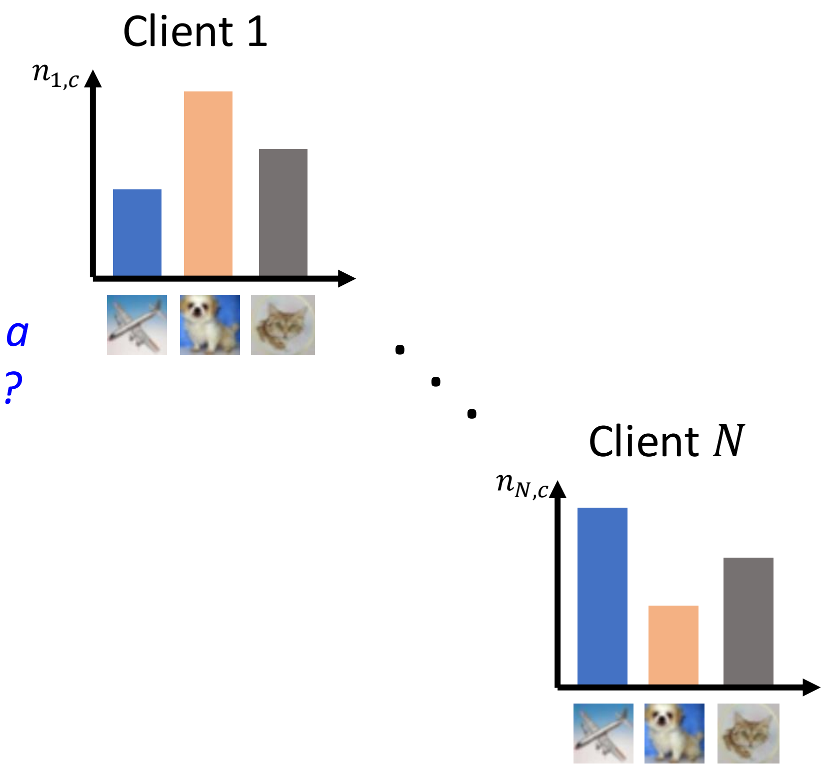
- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

- Summary

FedSLD – Background and motivation

- FedAvg assumption
  - Weighted sum of local empirical risks
  - Weights are often $n_i / \sum_j n_j$
  - Assumes knowledge of <u>number of samples</u>

Research question 1: *How can we leverage other sharable information to design a novel global FL algorithm for medical FL to mitigate the data heterogeneity issue?*

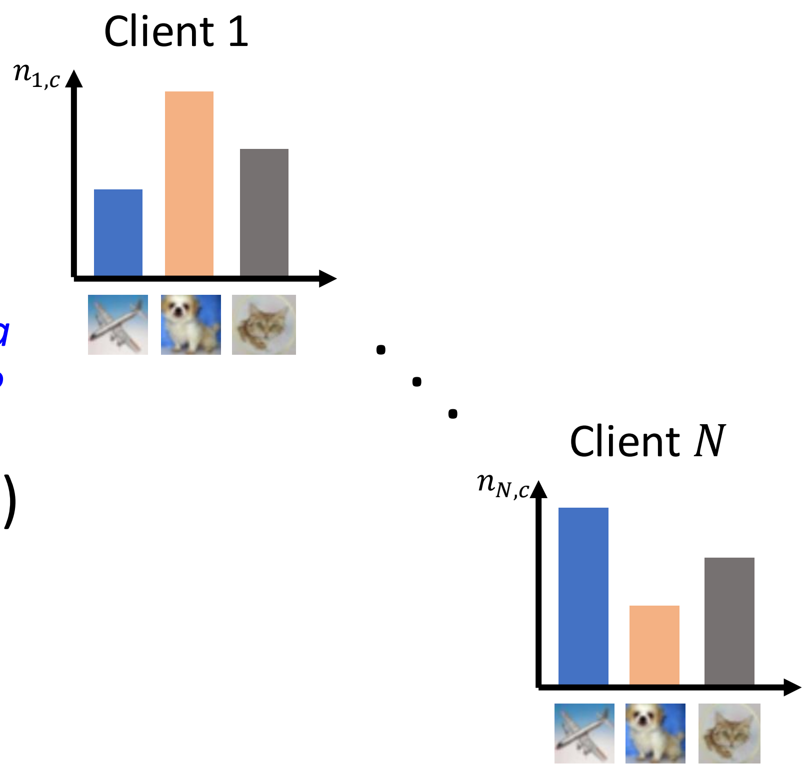


FedSLD – Background and motivation

- FedAvg assumption
  - Weighted sum of local empirical risks
  - Weights are often $n_i / \sum_j n_j$
  - Assumes knowledge of <u>number of samples</u>

Research question 1: *How can we leverage other sharable information to design a novel global FL algorithm for medical FL to mitigate the data heterogeneity issue?*

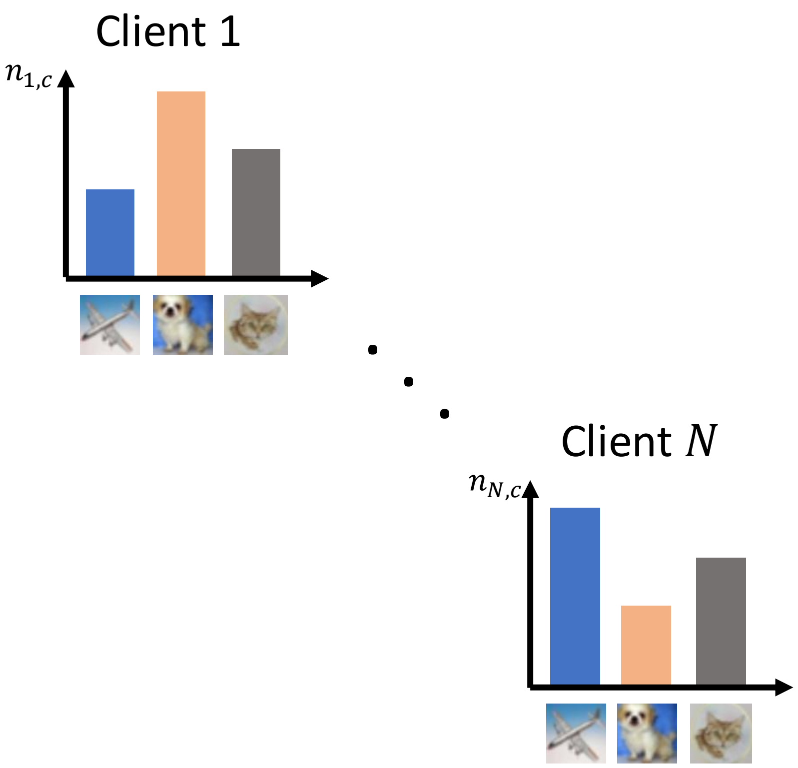
- **Fed**erated Learning with **S**hared **L**abel **D**istribution (**FedSLD**)
  - Leverages information and statistics regarding the local datasets
  - Assumes knowing <u>number of samples in each class</u>
    - This assumption usually holds true for medical cross-silo FL
    - Estimate of label distribution



FedSLD – Method

- Estimation of label distribution
  - Non-IID: $\mathcal{P}_i(x, y) \neq \mathcal{P}_j(x, y)$
  - By Bayes' theorem, $\mathcal{P}_i(x|y)\mathcal{P}_i(y) \neq \mathcal{P}_j(x|y)\mathcal{P}_j(y)$
  - Here, we only consider different $\mathcal{P}_i(y) \neq \mathcal{P}_j(y)$
  - Aggregate knowledge of #samples in each class, estimate $\mathcal{P}(y)$ by

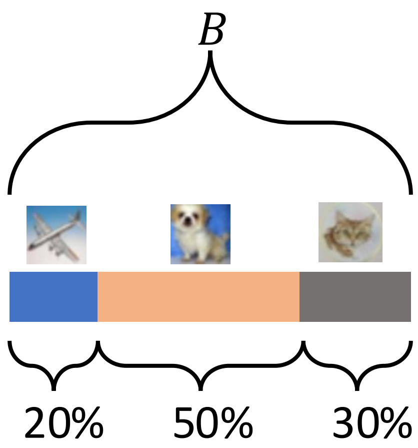
$$\tilde{\mathcal{P}}(y = c) = \frac{\sum_{i=1}^{N} n_{i,c}}{\sum_{i=1}^{N} n_i}$$



FedSLD – Method

- Compute the percentage of each class in each mini-batch
  - During local update, given a batch of data $\{(x_k, y_k)\}_{k=1}^{B}$ with B data samples, compute

$$p_b(y = c) = \frac{\sum_{k=1}^{B} [\![y_k = c]\!]}{B}$$



FedSLD – Method

- Weigh each data samples' contribution to the loss based on
  - The estimation of the prior of each class
  - The percentage of each class in each mini-batch

- Final loss of the mini-batch

$$\mathcal{L}_b(\{(x_k, y_k)\}_{k=1}^{B}) = -\sum_{k=1}^{B}\left(\boxed{\frac{\tilde{\mathcal{P}}(y = y_k)}{p_b(y = y_k)}} \cdot \sum_{c=1}^{C} y_{k,c} \log\big(f_i(x_k)\big)_c\right)$$

- Aggregate the model at the end of each training round as in FedAvg

Algorithm 1 FedSLD.

Input: Initialized model parameter weights w^0, number of clients N, number of local epochs E, batch size B, is the batch size, learning rate η, number of rounds R.

$\forall i \in [N], c \in [C]$, acquire $n_{i,c}$, client i's numbers of samples of each class c.
$\forall c \in [C]$, $\tilde{\mathcal{P}}(y = c) = \frac{\sum_{i=1}^{N} n_{i,c}}{\sum_{i=1}^{N} n_i}$ // compute estimated prior label distribution.
for $r \leftarrow 1, 2, ..., R$ **do**
  $\forall i \in [N]$ $w_i^r = w^{r-1}$ // broadcast model parameters.
  for $i \leftarrow 1, 2, ..., N$ **in parallel do**
    for $\{x_k, y_k\}_{k=1}^{B}$ **in all minibatches do**
      $\forall c$, $p_b(y = c) \leftarrow \sum_{k=1}^{B} [\![y_k = c]\!] / B$
      Compute loss $\mathcal{L}_b$ by Equation 3.
      $w_i^r \leftarrow w_i^r - \eta \nabla_w \mathcal{L}_b$
    end for
  end for
  $w^r = \sum_{i=1}^{N} \frac{n_i}{n} w_I^r$ // aggregate model updates
end for
return w^R

FedSLD – Experiments & results

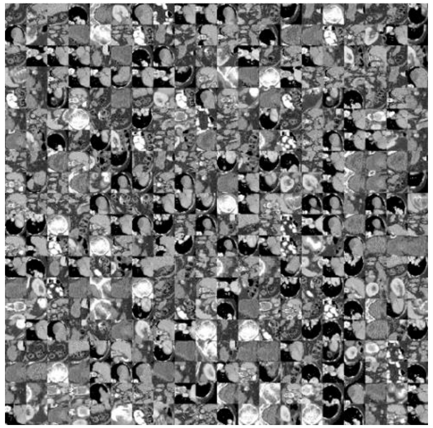
- Datasets

  - Two benchmark datasets
    - MNIST
    - CIFAR10

  - Two medical imaging datasets from MedMNIST collection
    - OrganMNIST (axial) (11-class liver tumor images)
    - PathMNIST (9-class colorectal cancer images)

PathMNIST

OraganMNIST (axial)



FedSLD – Experiments & results

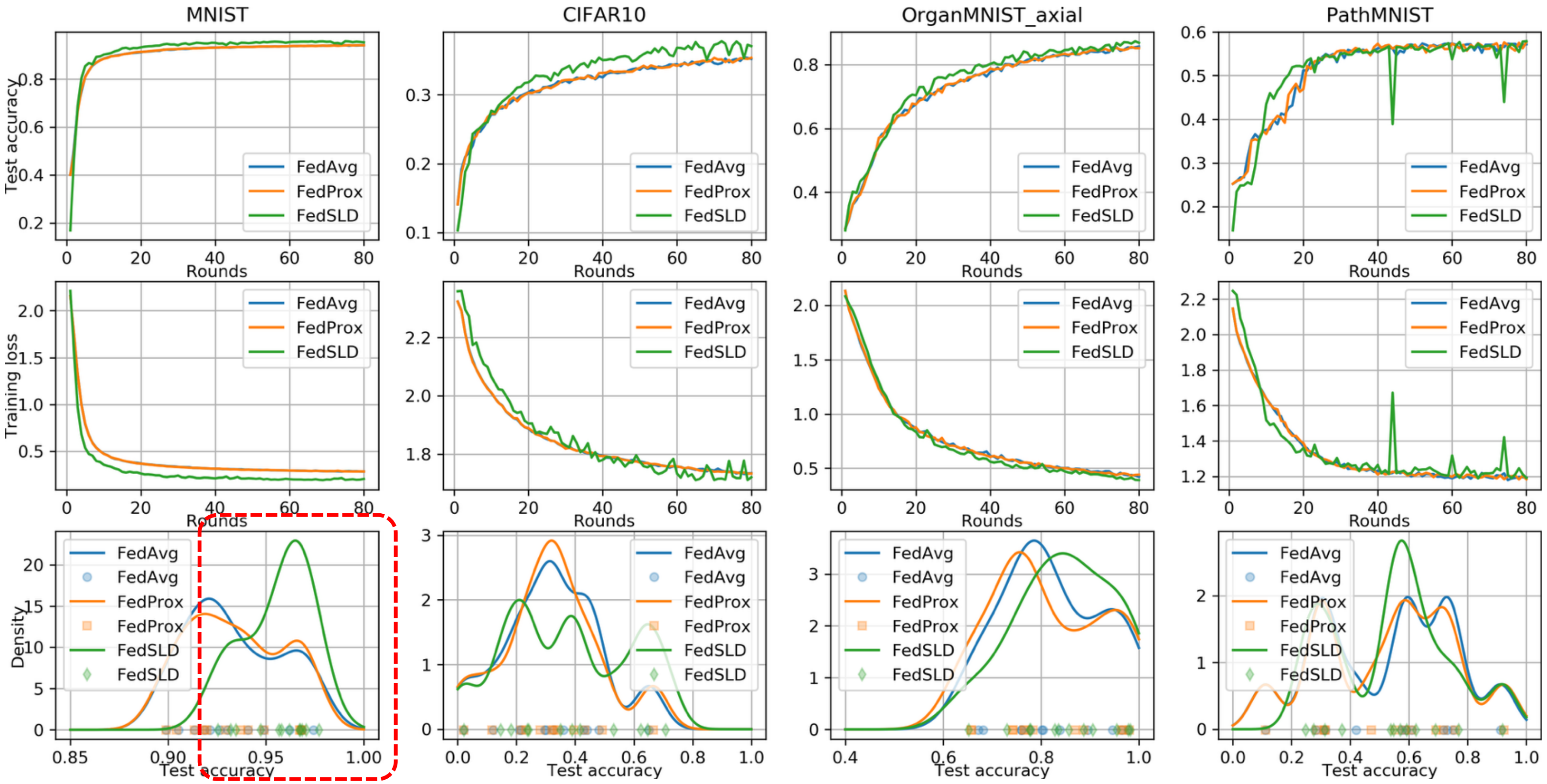
- Two non-IID settings
  - Pathological non-IID (12 clients)
    - Randomly select 2 classes for each client
    - In each class, assign a random number of images
  - Practical non-IID (12 clients)
    - Randomly partition each class of the dataset into 12 shards (10 x 1%, 1 x 10%, 1 x 80%)
    - Randomly assign one shard from each class to each client
    - A simulation that is closer to real-world medical applications
- Compared baselines
  - FedAvg
  - FedProx

FedSLD – Experiments & results

- Practical non-IID results

Mean personalized acc. / Combined test set (global) acc.

BMCTA/BTA	MNIST	CIFAR10	Organ-MNIST	Path-MNIST
FedAvg	93.41/94.15	32.07/35.46	82.32/85.69	52.70/57.38
FedProx	93.45/94.20	31.98/35.38	81.53/85.54	52.77/57.72
FedSLD (Ours)	**95.56/95.85**	**37.48/37.79**	**84.75/84.75**	**53.87/57.90**



Overview

- Federated learning: introduction

- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

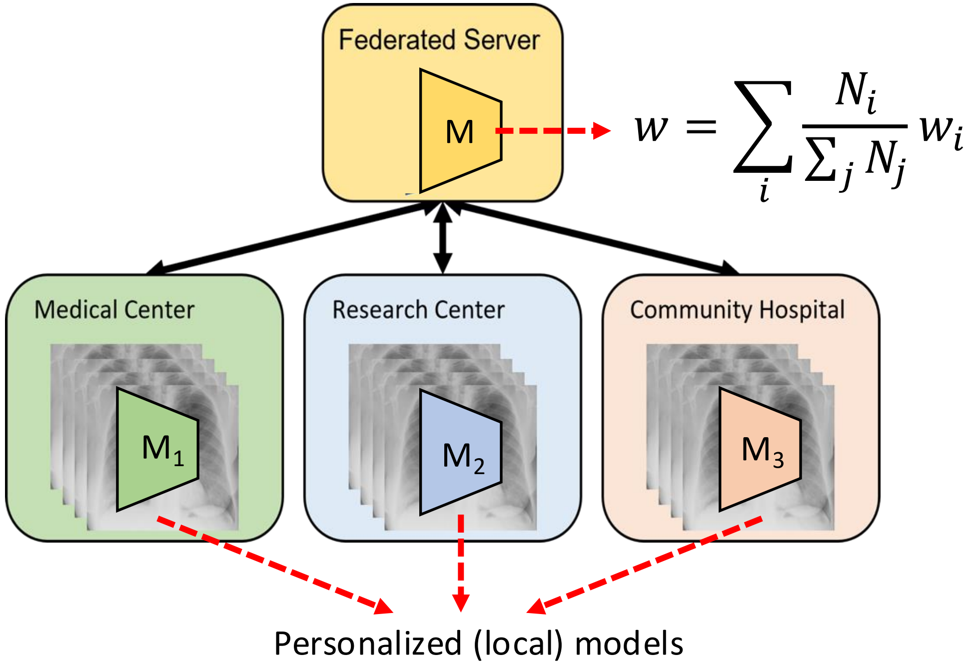
- Summary

PFL
IJCAI '22

APPLE – Background and motivation

- FedAvg aggregation
  - $w = \sum_i p_i w_i$
  - $p_i = N_i / \sum_j N_j$, aggregation weights are fixed
- Most existing FL/PFL methods
  - Use FedAvg-like aggregation
  - Training is either global or personalized

Research question 2: *How can we develop an **adaptive** aggregation strategy that optimally weighs different clients' contributions for each participant, while maintaining **a flexible balance** between global collaboration and local personalization objectives in cross-silo federated learning?*



APPLE – Background and motivation

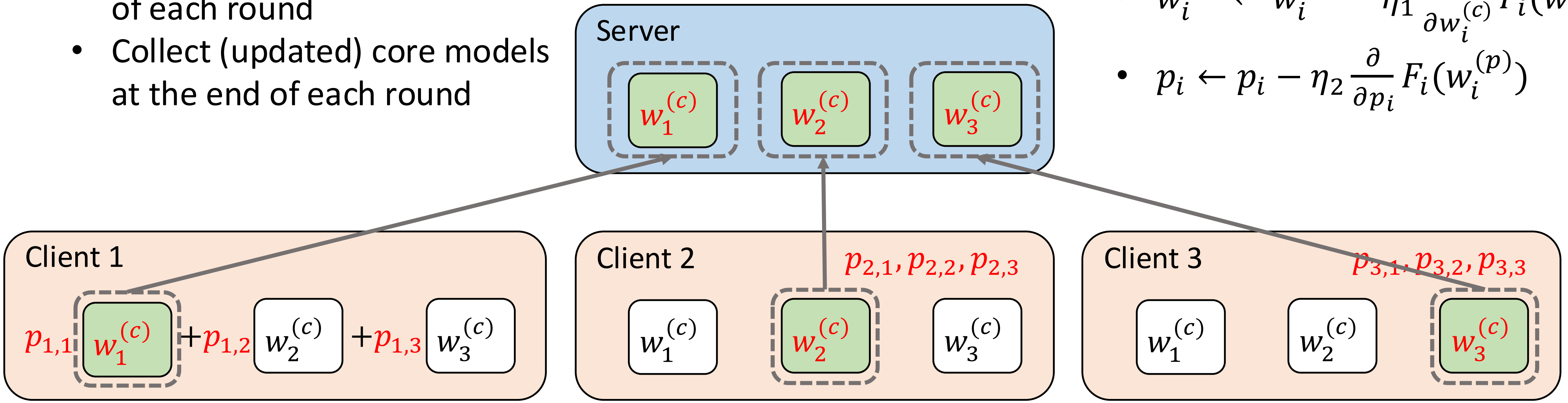
- ***A**da**p**tive **P**ersona**l**ized Cross-Silo F**e**derated Learning* (**APPLE**)
- The model of a client
  - **Personalized model $\boldsymbol{w_i^{(p)}}$**: used to do inference on client i
  - **Core model $\boldsymbol{w_i^{(c)}}$**: a constructing part of personalized model on client i, server also maintains core models from every client
  - $w_i^{(p)} = \sum_{j=1}^{N} p_{i,j} w_j^{(c)}$
  - **Directed relationship (DR) vector $\boldsymbol{p_i}$**: learnable weights (coefficients for core models) on client i, always kept locally

Client 1

$$w_1^{(p)} = p_{1,1} \boxed{w_1^{(c)}} + p_{1,2} \boxed{w_2^{(c)}} + p_{1,3} \boxed{w_3^{(c)}}$$

APPLE – Method

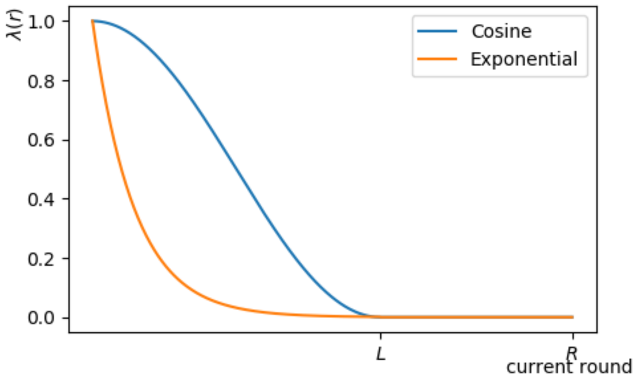
- Server
  - Broadcast core models to each client at the beginning of each round
  - Collect (updated) core models at the end of each round



- Local training
  - Clients' own core models and DR vectors are updated
  - $w_i^{(c)} \leftarrow w_i^{(c)} - \eta_1 \frac{\partial}{\partial w_i^{(c)}} F_i(w_i^{(p)})$
  - $p_i \leftarrow p_i - \eta_2 \frac{\partial}{\partial p_i} F_i(w_i^{(p)})$

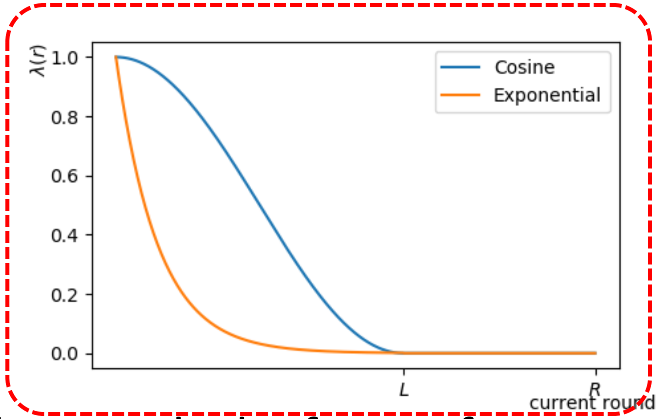
APPLE – Method

- Proximal Directed Relationships
  - Since downloaded core models are not trained from local empirical risk, training might be drawn to resembling individual learning (DR matrix drawn to identity matrix)
  - Penalize DR vector by a proximal term
  - $F_i\left(w_i^{(p)}\right) = \frac{1}{n_i}\sum_{\xi \in D_i^{tr}} \mathcal{L}\left(w_i^{(p)}; \xi\right) + \boxed{\lambda(r)\frac{\mu}{2}\|p_i - p_0\|_2^2}$
  - Prox-center $p_0 = [\frac{n_1}{n}, \dots, \frac{n_N}{n}]$
  - Loss scheduler $\lambda(r) \in [0,1]$: a decreasing function w.r.t. current round, controls the focus of training; μ: the peak value of the proximal term coefficient
  - Proximal term coefficient: ∞ → FedAvg; large → facilitate learning global high-level feature; small → concentrate on local empirical risk, learning the personalization



APPLE – Method

- Proximal Directed Relationships
  - Since downloaded core models are not trained from local empirical risk, training might be drawn to resembling individual learning (DR matrix drawn to identity matrix)
  - Penalize DR vector by a proximal term
  - $F_i\left(w_i^{(p)}\right) = \frac{1}{n_i}\sum_{\xi \in D_i^{tr}} \mathcal{L}\left(w_i^{(p)}; \xi\right) + \boxed{\lambda(r)}\frac{\mu}{2}\|p_i - p_0\|_2^2$
  - Prox-center $p_0 = [\frac{n_1}{n}, \ldots, \frac{n_N}{n}]$
  - Loss scheduler $\lambda(r) \in [0,1]$: a decreasing function w.r.t. current round, controls the focus of training; μ: the peak value of the proximal term coefficient
  - Proximal term coefficient: ∞ → FedAvg; large → facilitate learning global high-level feature; small → concentrate on local empirical risk, learning the personalization



APPLE – Experiments

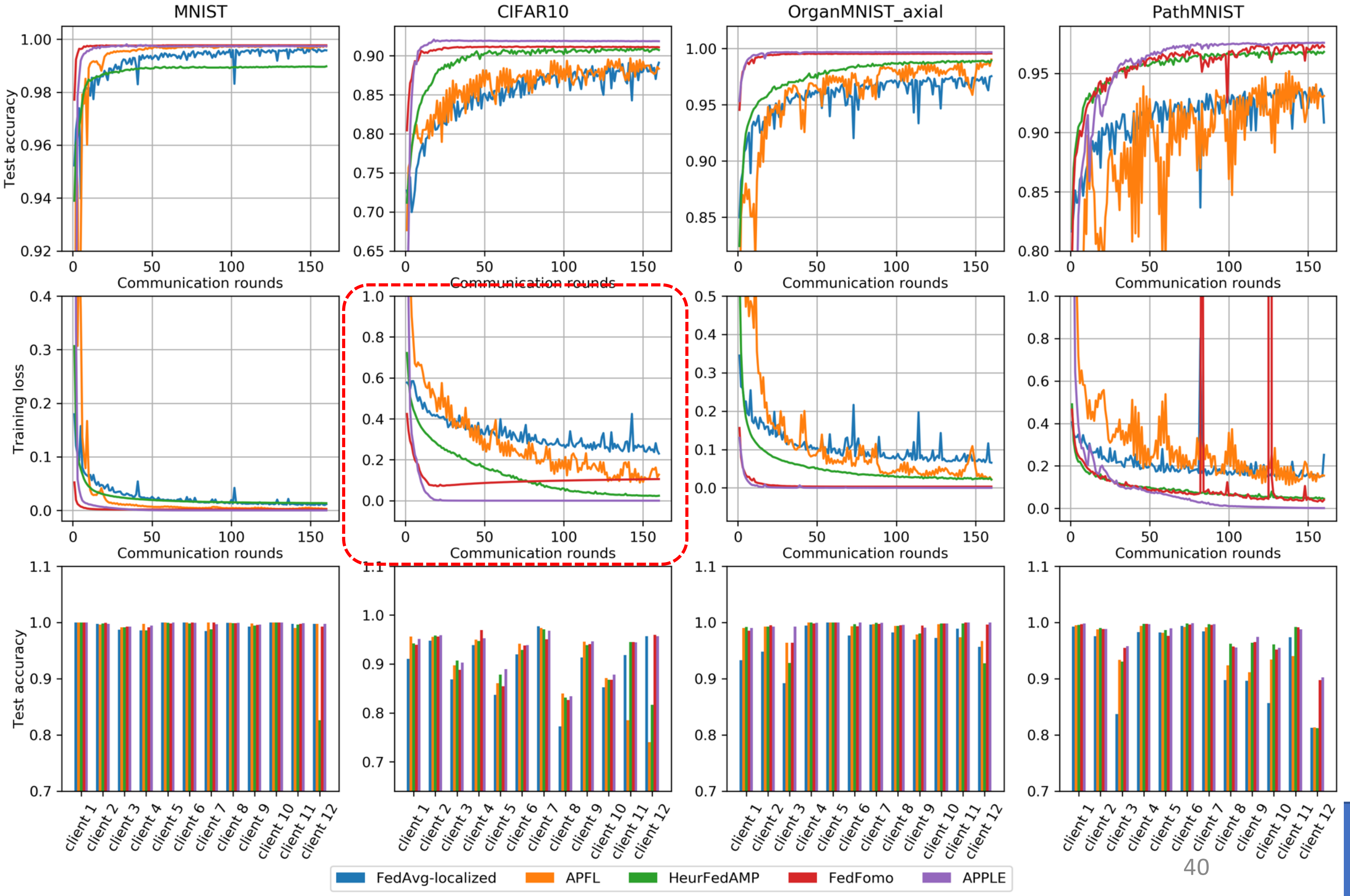
- Datasets
  - MNIST
  - CIFAR10
  - OraganMNIST (axial)
  - PathMNIST
- Two non-IID settings (same with FedSLD)
  - Pathological non-IID
  - Practical non-IID

- Compared baselines
  - Separate training
  - FedAvg (McMahan et al., 2017)
  - FedAvg-local
  - FedAvg-FT, FedProx-FT (Wang et al., 2019)
  - APFL (Deng et al., 2020)
  - HeurFedAMP (Huang et al., 2021)
  - FedFomo (Zhang et al., 2021)

APPLE – Results

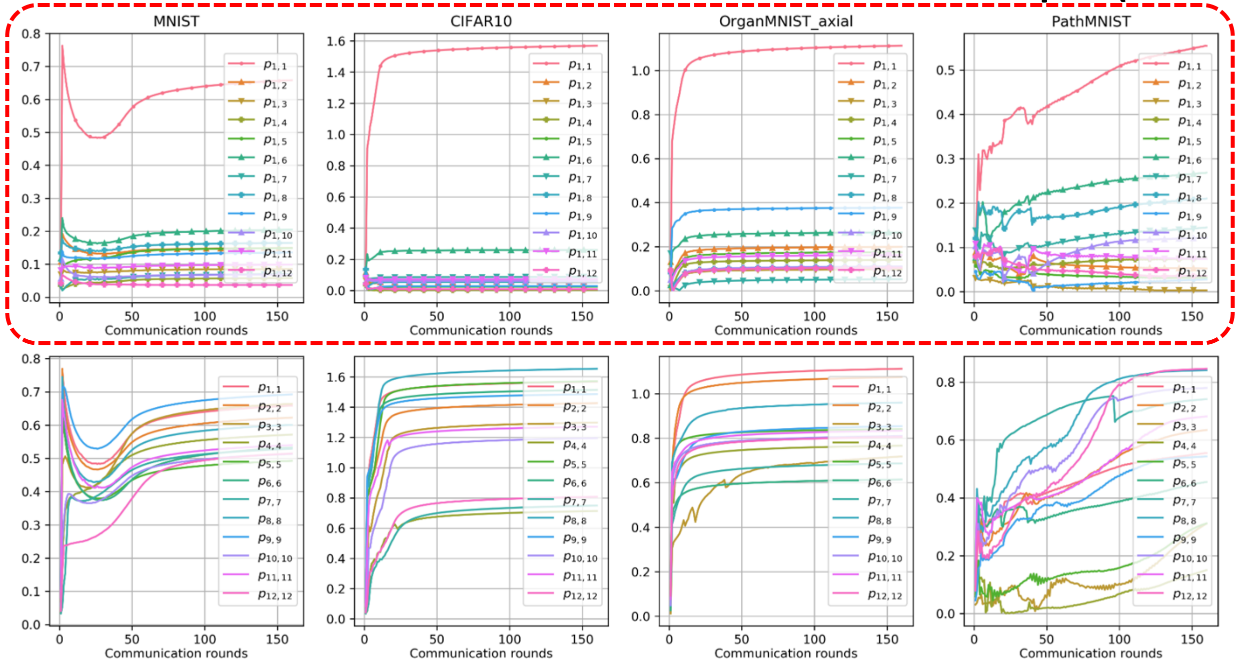
- Pathological non-IID

	Pathological non-IID			
	MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST
Separate	97.34	74.96	93.14	87.09
FedAvg	95.71	51.44	59.43	56.61
FedAvg-local	99.52	90.10	96.76	93.21
FedAvg-FT	99.43	90.49	97.03	92.31
FedProx-FT	99.43	90.49	97.03	92.38
APFL	99.75	89.30	98.72	94.98
HeurFedAMP	98.13	91.10	98.39	96.55
FedFomo	99.71	91.96	99.31	97.24
APPLE, $\mu = 0$	99.73	92.22	**99.66**	96.78
APPLE, $\mu \neq 0$	**99.77**	**92.68**	99.61	**97.51**

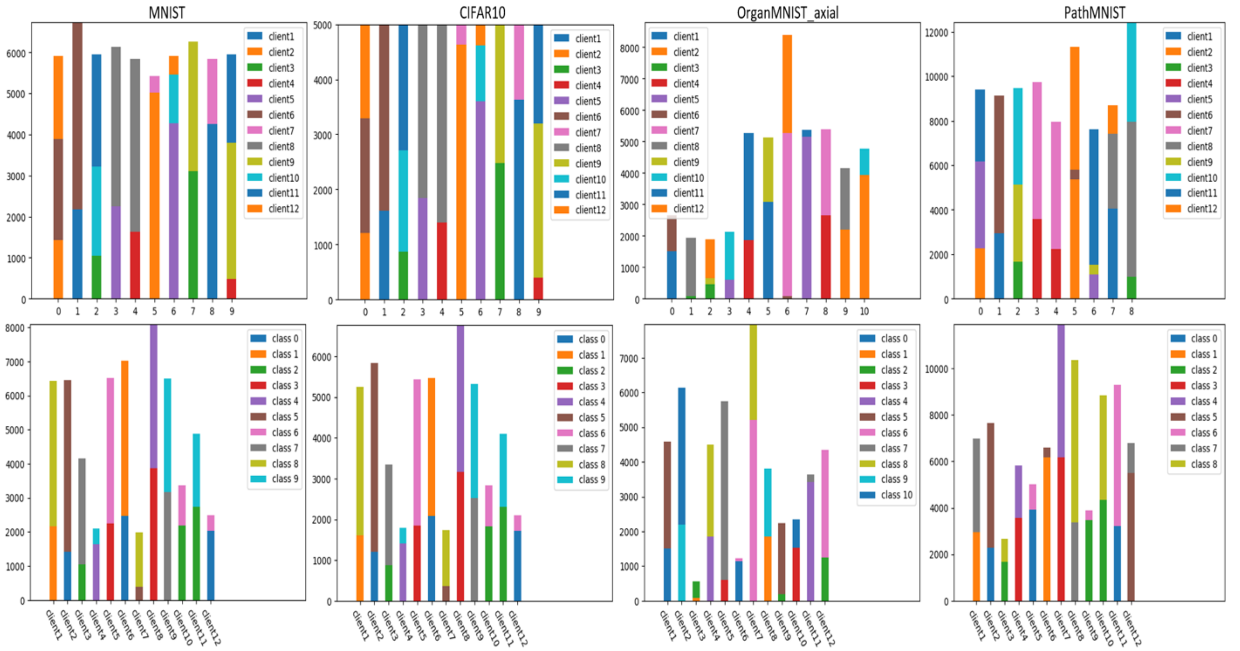


APPLE – Results

- Visualization of Directed Relationships (Pathological non-IID)



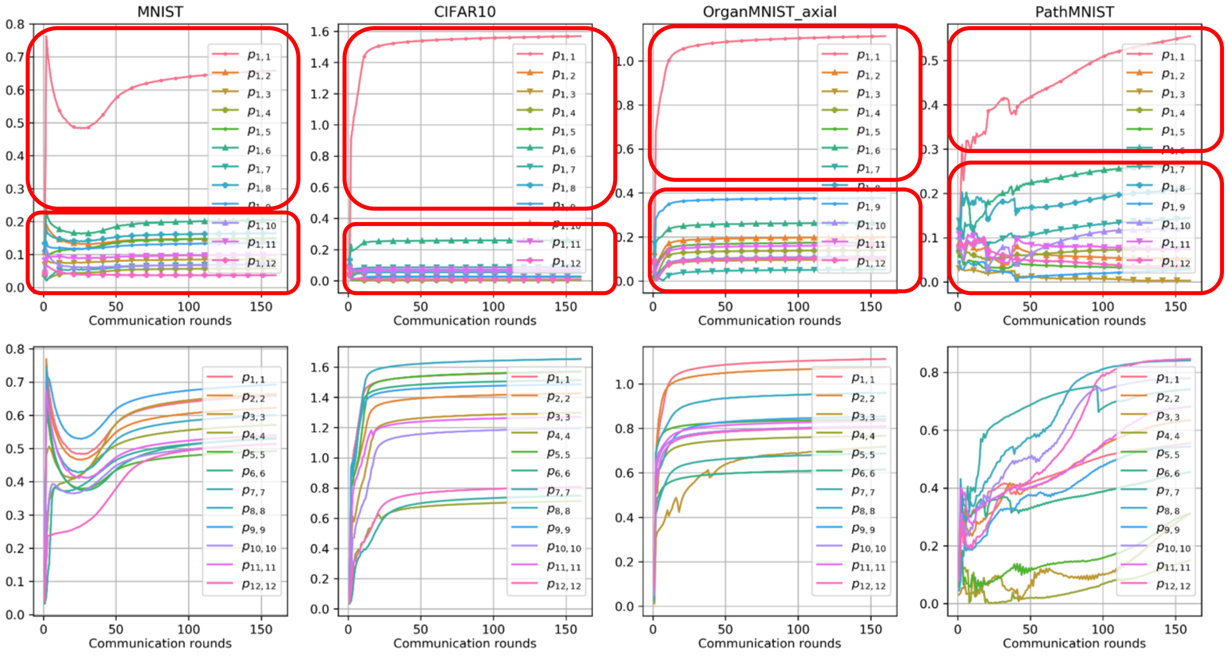
Visualization of DR



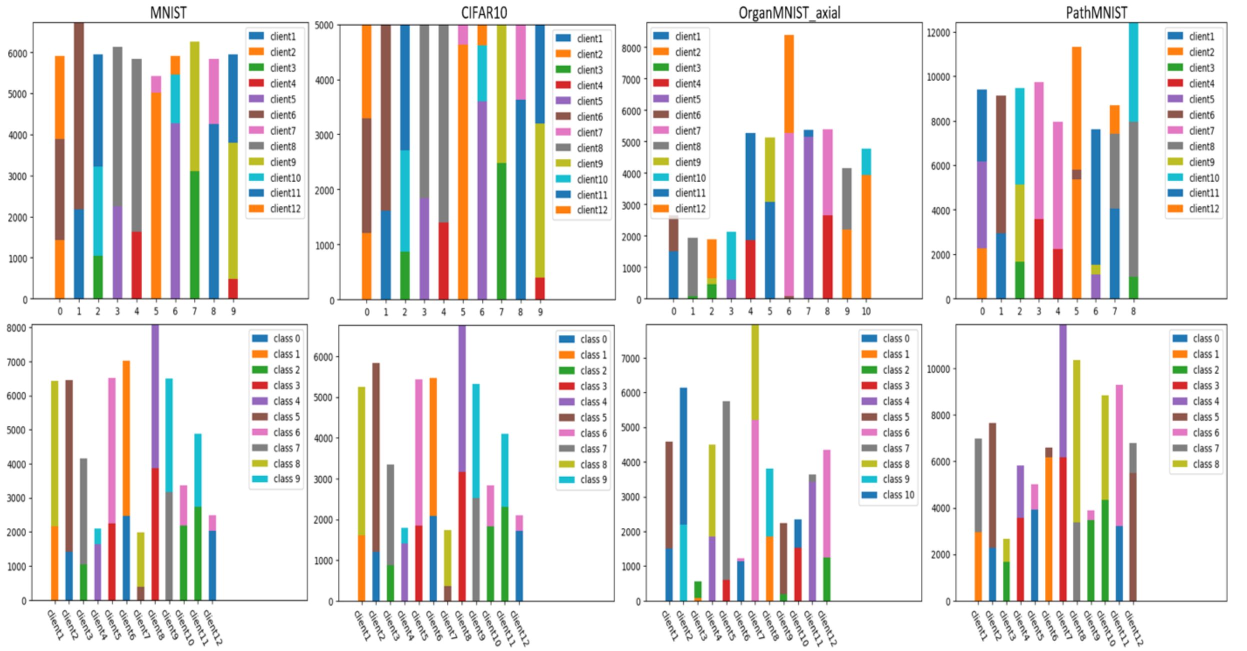
Data distribution

APPLE – Results

- Visualization of Directed Relationships (Pathological non-IID)



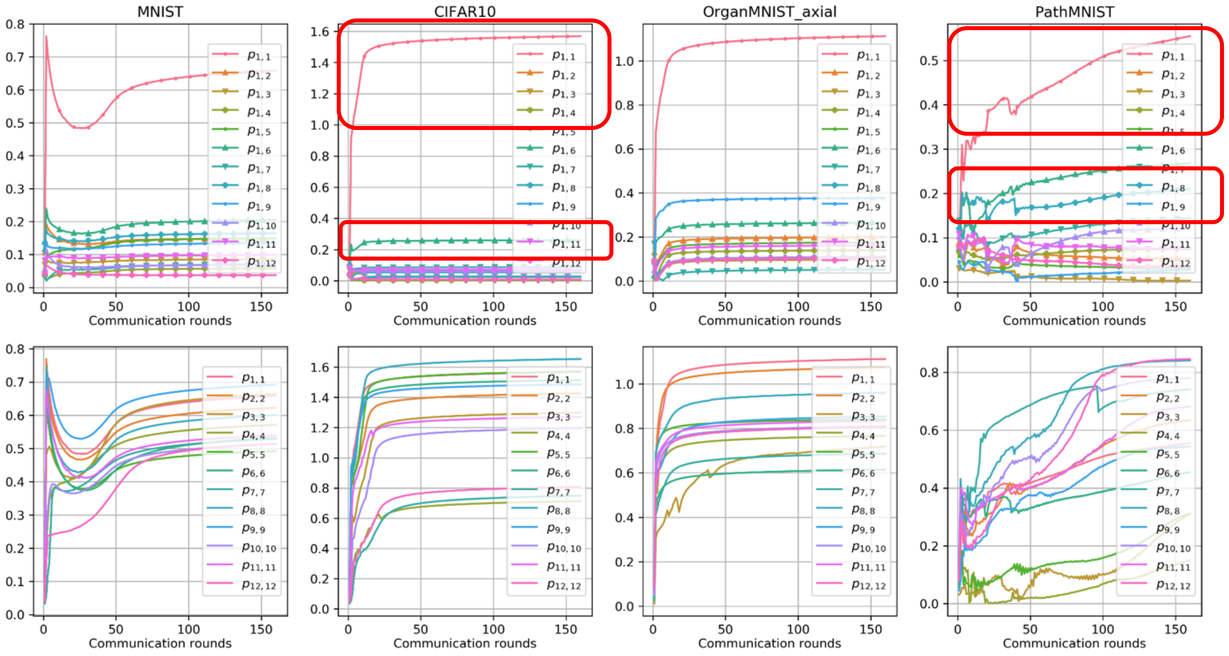
Visualization of DR



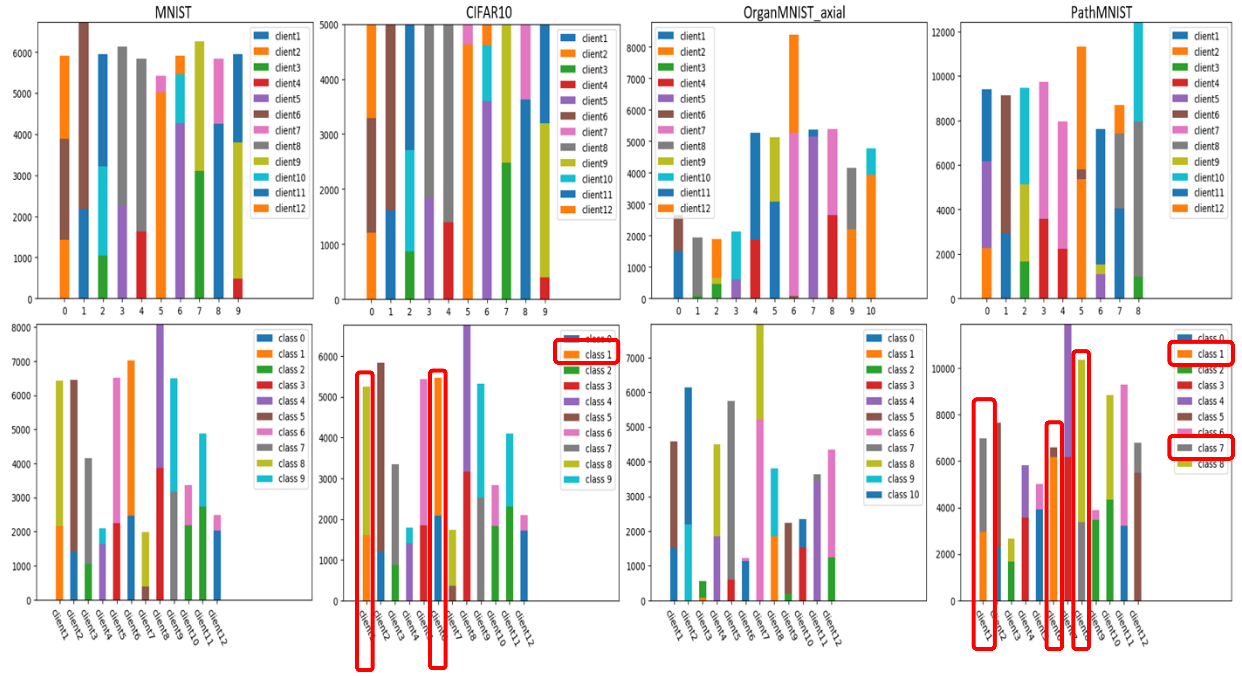
Data distribution

APPLE – Results

- Visualization of Directed Relationships (Pathological non-IID)



Visualization of DR

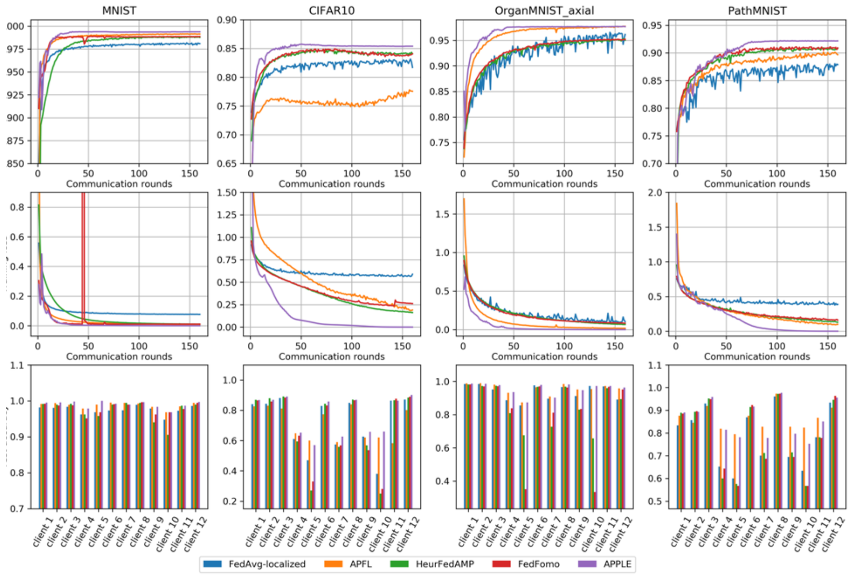


Data distribution

APPLE – Results

- Practical non-IID

	Practical non-IID			
	MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST
Separate	78.20	63.06	65.21	61.36
FedAvg	94.00	34.32	86.56	53.83
FedAvg-local	97.47	71.99	93.75	78.70
FedAvg-FT	97.66	72.08	94.13	78.69
FedProx-FT	97.66	72.08	94.13	78.69
APFL	98.80	71.19	95.53	86.35
HeurFedAMP	97.45	69.54	86.82	79.33
FedFomo	98.05	70.15	82.86	79.39
APPLE, $\mu = 0$	**99.00**	75.62	**95.70**	84.22
APPLE, $\mu \neq 0$	98.97	**77.41**	95.62	**86.39**



- Under limited bandwidth

		Pathological non-IID				Practical non-IID			
		MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST	MNIST	CIFAR10	Organ-MNIST (axial)	Path-MNIST
$M = 11$	FedFomo	99.71	91.96	99.31	97.24	98.05	70.15	82.86	79.39
	APPLE	99.73	92.22	99.66	96.78	99.00	75.62	95.70	84.22
$M = 7$	FedFomo	99.71	91.95	99.31	97.33	97.65	70.24	80.88	80.19
	APPLE	99.73	92.17	99.53	97.15	98.70	76.14	94.21	84.07
$M = 5$	FedFomo	99.71	91.94	99.31	97.40	97.47	70.44	82.83	79.62
	APPLE	99.72	92.28	99.48	97.17	98.45	75.63	94.49	85.46
$M = 2$	FedFomo	99.71	91.98	99.31	97.25	96.51	69.87	79.53	79.26
	APPLE	99.70	92.41	99.47	97.11	98.29	74.84	92.29	84.64
$M = 1$	FedFomo	99.71	91.95	99.31	97.15	91.54	69.93	78.37	75.17
	APPLE	99.66	92.31	99.59	96.29	98.52	73.03	93.55	83.35

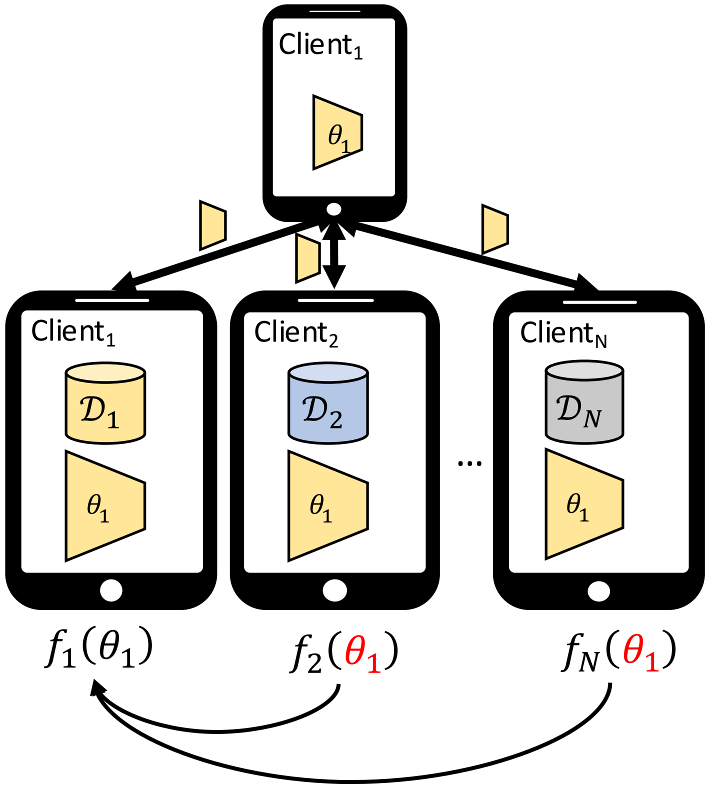
Overview

- Federated learning: introduction
- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)

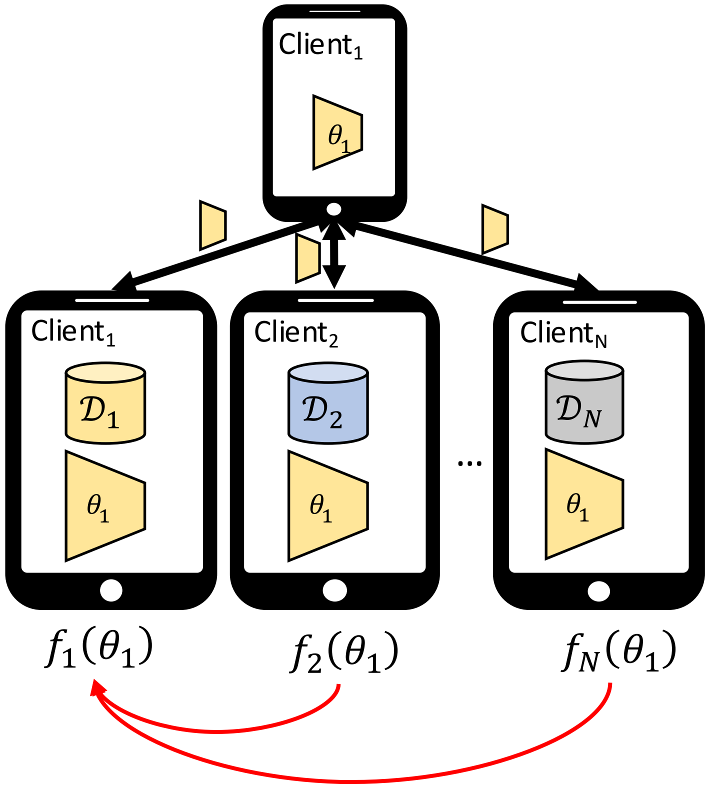
PFL
ICCV '23 oral

- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection
- Summary

PGFed – Background and motivation



PGFed – Background and motivation

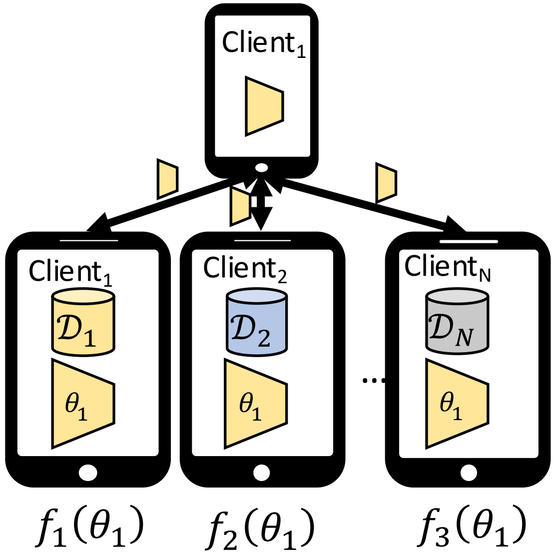


Explicit

collaborative knowledge transfer

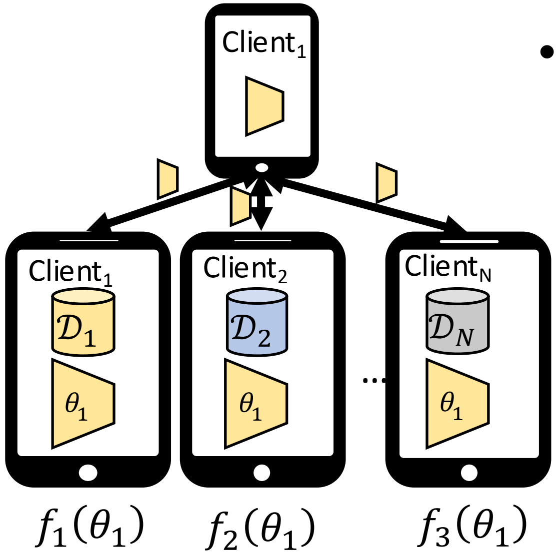
PGFed – Background and motivation

- Why explicit (especially for personalized model update)?
  - (**Explicitness**: Direct engagement of multiple clients’ empirical risks)
  - Intuition/motivation: facilitate the generalizability of θ_i directly by penalizing its performance over other clients’ empirical risks.

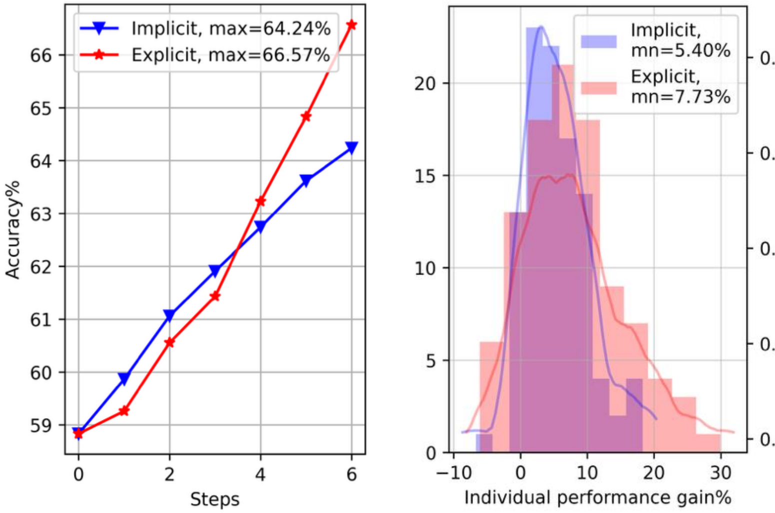


PGFed – Background and motivation

- Why explicit (especially for personalized model update)?
  - (**Explicitness**: Direct engagement of multiple clients' empirical risks)
  - Intuition/motivation: facilitate the generalizability of θ_i directly by penalizing its performance over other clients' empirical risks.

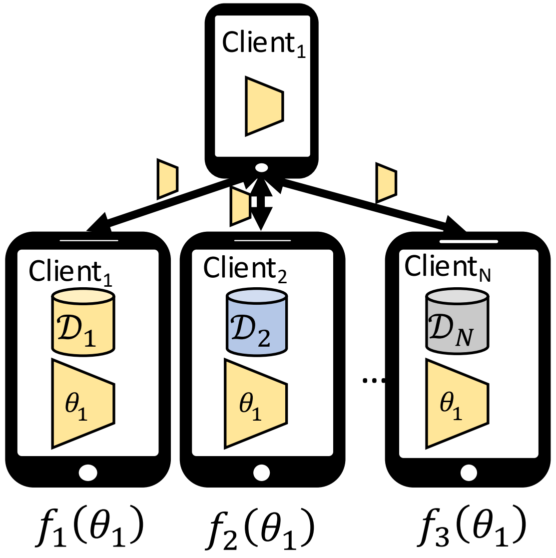


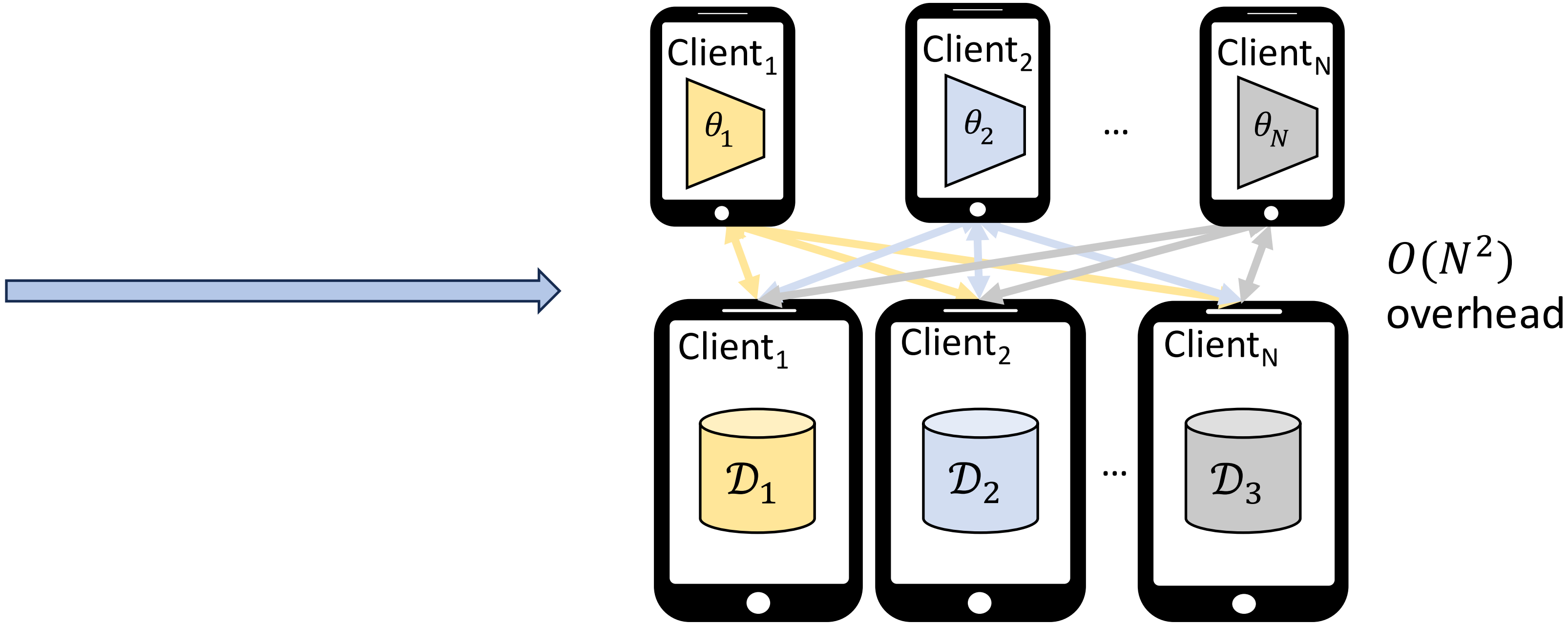
- Toy experiment on exemplar design
  - Cifar10, 100 heterogeneous clients
  - Explicit: $\boxed{F_i(\theta_i)} = f_i(\theta_i) + \boxed{\frac{\mu}{N-1}\sum_{j\neq i} f_j(\theta_i)}$
  - Implicit: $F_i(\theta_i) = f_i(\theta_i)$ (local model of FedAvg)



PGFed – Background and motivation

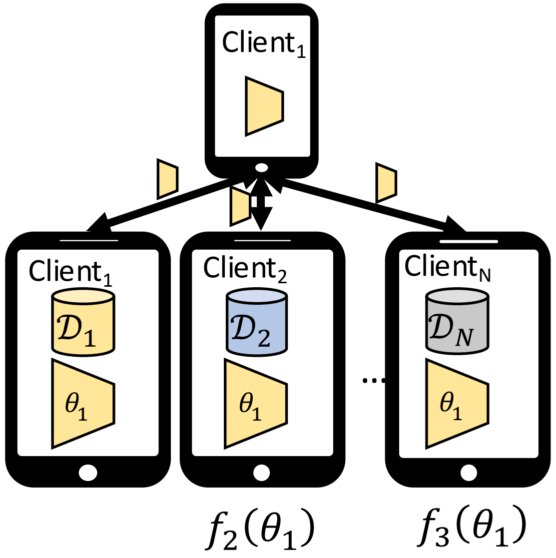
- Why explicit (especially for personalized model update)?
  - (**Explicitness**: Direct engagement of multiple clients' empirical risks)
  - Intuition/motivation: facilitate the generalizability of θ_i directly by penalizing its performance over other clients' empirical risks.



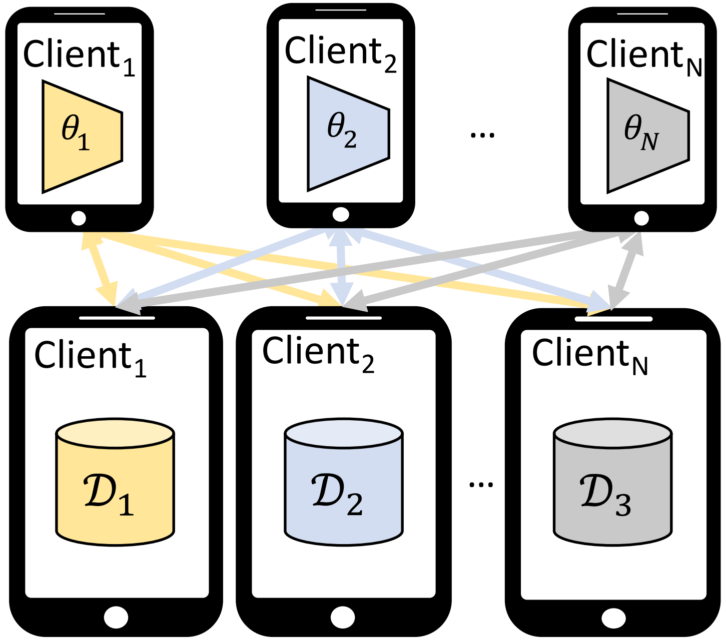


PGFed – Background and motivation

- Why explicit (especially for personalized model update)?
  - (**Explicitness**: Direct engagement of multiple clients' empirical risks)
  - Intuition/motivation: facilitate the generalizability of θ_i directly by penalizing its performance over other clients' empirical risks.



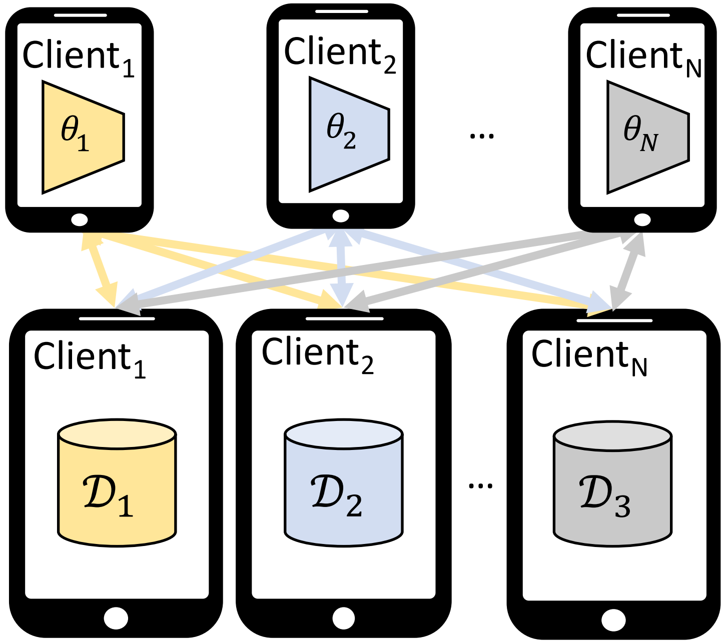
Research question 3: *How can we design an **explicit** PFL framework to further boost the model performance with **linear communication** complexity that remains practical for both cross-silo and cross-device federated learning scenarios?*

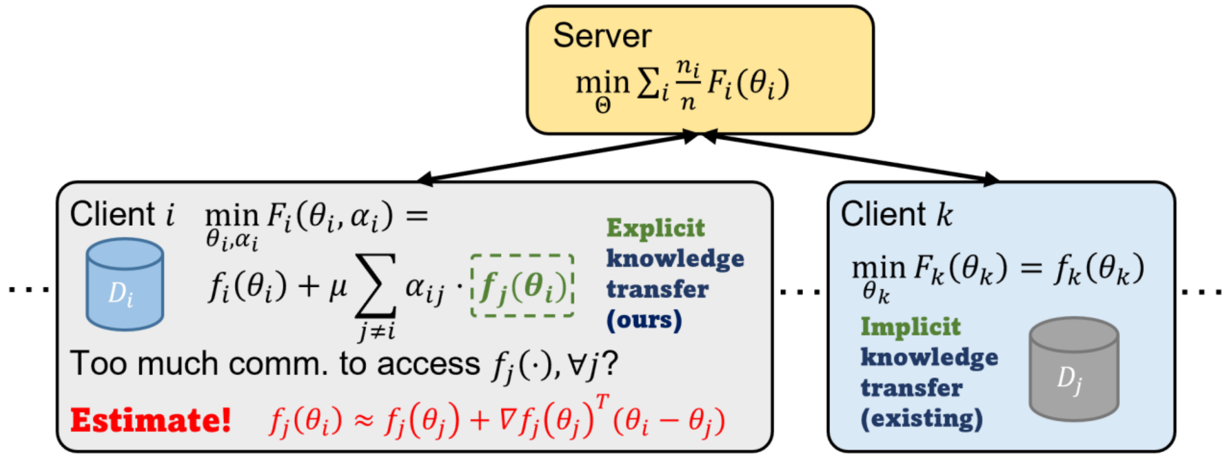


PGFed – Background and motivation

- Difficulty to achieve explicitness
  - $O(N^2)$ communication overhead ⟹ ✓ Estimate $f_j(\theta_i) \approx f_j(\theta_j) + \nabla f_j(\theta_j)^T(\theta_i - \theta_j)$, $O(N^2) \rightarrow O(N)$
  - Proper coefficient for each non-local risk ⟹ ✓ Use adaptive coefficient $\alpha_{ij} \forall i,j \in [N]$

- Proposed solution: **Personalized Global FL (PGFed)**





PGFed – Method

- Objectives of Personalized Global Federated Learning (**PGFed**)
  - Global objective: $\min_{\boldsymbol{\Theta}, \boldsymbol{A}} F(\boldsymbol{\Theta}, \boldsymbol{A}) = \min_{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N} \sum_{i=1}^{N} p_i F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i)$
  - Local objective: $F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = f_i(\boldsymbol{\theta}_i) + \mu \sum_{j \in [N]} \alpha_{ij} \boxed{f_j(\boldsymbol{\theta}_i)}$
  - Plugging $\boxed{f_j(\theta_i) \approx f_j(\theta_j) + \nabla f_j(\theta_j)^T (\theta_i - \theta_j)}$ into Local objective, we have

$$F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) \approx f_i(\boldsymbol{\theta}_i) + \mathcal{R}_{aug}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i)$$

$$\mathcal{R}_{aug}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = \mu \sum_{j \in [N]} \alpha_{ij} \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \right)$$

$$\mathcal{R}_{aug}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = \mu \sum_{j\in[N]} \alpha_{ij} \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \right)$$

PGFed – Method

- Gradient-based update
  - W.r.t θ_i:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_i} F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) &= \nabla_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i) + \nabla_{\boldsymbol{\theta}_i} \mathcal{R}_{aug}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) \\ &= \nabla_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i) + \underbrace{\mu \sum_{j\in[N]} \alpha_{ij} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)}_{\tilde{\boldsymbol{g}}_{[N]}} . \end{aligned}$$

    - $\tilde{g}_{[N]}$ can be computed by the server with:
      - Client i uploading α_i
      - Client j uploading local gradient

$\mathcal{R}_{aug}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = \mu \sum_{j\in[N]} \alpha_{ij} \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\right)$

PGFed – Method

- Gradient-based update
  - W.r.t α_{ij}:

$$\begin{aligned}\nabla_{\alpha_{ij}} F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) &= \mu \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\right) \\ &= \underbrace{\mu \left(f_j(\boldsymbol{\theta}_j) - \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T \boldsymbol{\theta}_j\right)}_{g_\alpha^{(1)}} + \underbrace{\mu \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T \boldsymbol{\theta}_i}_{g_\alpha^{(2)}} .\end{aligned}$$

    - $g_\alpha^{(1)}$ (a scalar) can be computed and uploaded by the client j
    - $g_\alpha^{(2)}$ (exact value needs to transmit all gradients to client i (takes $O(N^2)$ comm.))
      - Estimate: $g_\alpha^{(2)} \approx \bar{\boldsymbol{g}}_{[N]}^T \boldsymbol{\theta}_i = \frac{\mu}{N} \left(\sum_{j\in[N]} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)\right)^T \boldsymbol{\theta}_i$
        - Client j uploading local gradient

PGFed – Method

- To accommodate to M selected clients per round: $[N] \rightarrow S_t$ (selected set of clients in round t)

$$\tilde{\boldsymbol{g}}_{\mathcal{S}_t} = \mu \sum_{j \in \mathcal{S}_t} \alpha_{ij} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \qquad \bar{\boldsymbol{g}}_{\mathcal{S}_t} = \frac{\mu}{M} \left(\sum_{j \in \mathcal{S}_t} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \right)$$

- To keep information from clients selected in previous round, use momentum (PGFedMo)

$$\tilde{\boldsymbol{g}}^i_{\mathcal{S}_t} = (1-\beta)\tilde{\boldsymbol{g}}^i_{\mathcal{S}_t}(\text{downloaded}) + \beta \tilde{\boldsymbol{g}}^i_{\mathcal{S}_t}(\text{previous})$$

Algorithm 1 `PGFed` and `PGFedMo`

Input: N clients, learning rates η_1, η_2, number of rounds T, coefficient μ(, momentum β for `PGFedMo`)
Output: Personalized models $\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_N^T$.

ServerExecute:

Initialize $\alpha_{ij} \leftarrow 1/M \ \forall i, j \in [N]$, global model $\boldsymbol{\theta}^0_{glob}$
$\mathbf{A}[i] \leftarrow \boldsymbol{\alpha}_i \ \forall i \in [N]$
for $t \leftarrow 1, 2, ..., T$ **do**
  Select a subset of M clients, $\mathcal{S}_t$
  $g_t^{(1)} \leftarrow \{\}; \nabla_t \leftarrow \{\}$ // built for next round
  for $i \in \mathcal{S}_t$ **in parallel do**
    if t=1 **then**
      $\boldsymbol{\theta}_i^t, g_\alpha^{(1)}, \nabla f(\boldsymbol{\theta}_i^t), \boldsymbol{\alpha}_i \leftarrow$ **ClientUpdate**($\boldsymbol{\theta}^{t-1}_{glob}, t$)
    else
      $\tilde{\boldsymbol{g}}_{\mathcal{S}_{t-1}} \leftarrow \mu \sum_{j \in \mathcal{S}_{t-1}} \alpha_{ij} \nabla_{t-1}[j]$
      $\bar{\boldsymbol{g}}_{\mathcal{S}_{t-1}} \leftarrow \frac{\mu}{M} \sum_{j \in \mathcal{S}_{t-1}} \nabla_{t-1}[j]$
      $\boldsymbol{\theta}_i^t, g_\alpha^{(1)}, \nabla f(\boldsymbol{\theta}_i^t), \boldsymbol{\alpha}_i \leftarrow$ **ClientUpdate**($\boldsymbol{\theta}^{t-1}_{glob}, t, \tilde{\boldsymbol{g}}_{\mathcal{S}_{t-1}}, \bar{\boldsymbol{g}}_{\mathcal{S}_{t-1}}, g_{t-1}^{(1)}$)
    end if
    // the next line records the values for next round
    $\mathbf{A}[i] \leftarrow \boldsymbol{\alpha}_i; g_t^{(1)}[i] \leftarrow g_\alpha^{(1)}; \nabla_t[i] \leftarrow \nabla f(\boldsymbol{\theta}_i^t)$
    $\boldsymbol{\theta}^t_{glob} \leftarrow \sum_{i \in \mathcal{S}_t} p_i \boldsymbol{\theta}_i^t$
  end for
  for $i \in ([N] - \mathcal{S}_t)$ **in parallel do**
    $\boldsymbol{\theta}_i^t \leftarrow \boldsymbol{\theta}_i^{t-1}; \tilde{\boldsymbol{g}}_i^t \leftarrow \tilde{\boldsymbol{g}}_i^{t-1}$
  end for
end for
return $\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_N^T$

ClientUpdate($\boldsymbol{\theta}^{t-1}_{global}, t$ (, $\tilde{\boldsymbol{g}}, \bar{\boldsymbol{g}}, g_{t-1}^{(1)}$)):

if t=1 **then**
  $\boldsymbol{\theta}_i^t \leftarrow$ **ClientUpdate**($\boldsymbol{\theta}^{t-1}_{global}, \eta_1$) as in `FedAvg`
else
  $\boldsymbol{\theta}_i^t \leftarrow \boldsymbol{\theta}^{t-1}_{global}$
  $\tilde{\boldsymbol{g}}_i^t \leftarrow \tilde{\boldsymbol{g}}$ // without momentum
  $\tilde{\boldsymbol{g}}_i^t \leftarrow (1-\beta)\tilde{\boldsymbol{g}} + \beta \tilde{\boldsymbol{g}}_i^{t-1}$ // with momentum
  for Batch of data $\mathcal{B} \in \mathcal{D}_i$ **do**
    $\boldsymbol{\theta}_i^t \leftarrow \boldsymbol{\theta}_i^t - \eta_1(\nabla f(\boldsymbol{\theta}_i^t, \mathcal{B}) + \tilde{\boldsymbol{g}}_t^i)$
    $g^{(2)} = \bar{\boldsymbol{g}}^T \boldsymbol{\theta}_i$
    $\forall j \in g_{t-1}^{(1)}: \ \alpha_{ij} \leftarrow \alpha_{ij} - \eta_2(g_{t-1}^{(1)}[j] + g^{(2)})$
  end for
end if
$g_\alpha^{(1)} \leftarrow \mu \left(f(\boldsymbol{\theta}_i^t) - \nabla f(\boldsymbol{\theta}_i^t)^T \boldsymbol{\theta}_i^t \right)$ // for next round
return $\boldsymbol{\theta}_i^t, g_\alpha^{(1)}, \nabla f(\boldsymbol{\theta}_i^t), \boldsymbol{\alpha}_i$

PGFed – Experiments & results

- Settings
  - Datasets: CIFAR10, CIFAR100, OrganMNIST, Office-home
  - Partition
    - CIFAR10/100:(Dir($\alpha = 0.3$)), 25, 50, 100 clients, 25% sample rate
    - OrganMNIST: 25 clients, Dir($\alpha = 1.0$), 50% sample rate
      50, 100 clients, Dir($\alpha = 0.3$), 25% sample rate
    - Office-home: 5 clients/domain x 4 domains, Dir($\alpha = 0.3$), 25% sample rate
  - Metric: mean personalized test accuracy
  - Compared methods
    - Local
    - FedAvg
    - FedDyn
    - pFedMe
    - FedFomo
    - APFL
    - FedReP
    - LG-FedAvg
    - FedPer
    - Per-FedAvg
    - FedRoD
    - FedBABU

Heterogeneous partition of a dataset based on Dirichlet distribution:

- $\alpha = \infty$ → homogeneous
- α = 0.3/0.5/1.0 → very heterogeneous, with 1.0 slightly balanced (tend to have lower acc.)
- $\alpha = 0$ → one class per client

PGFed – Experiments & results

- Performance on CIFAR10 & CIFAR100

	CIFAR10 25 clients	CIFAR10 50 clients	CIFAR10 100 clients	CIFAR100 25 clients	CIFAR100 50 clients	CIFAR100 100 clients
Local	72.40±0.45	70.28±0.38	67.39±0.20	32.74±0.08	26.05±0.34	23.06±0.47
FedAvg [26]	65.07±0.25	64.41±0.66	63.19±0.46	28.48±0.59	26.06±0.65	25.58±0.80
FedDyn [1]	67.31±0.36	65.02±0.91	62.49±0.06	34.17±0.43	27.06±0.18	23.88±0.36
pFedMe [35]	70.60±0.23	68.92±0.35	66.40±0.04	27.97±0.24	23.82±0.06	22.35±0.03
FedFomo [43]	72.33±0.03	72.17±0.48	70.86±0.27	32.15±0.61	25.90±1.17	24.48±0.44
APFL [6]	77.03±0.26	77.36±0.18	76.29±0.13	39.16±0.93	35.15±0.65	33.86±0.60
FedRep [5]	76.85±0.44	76.03±0.17	72.30±0.52	33.43±0.80	26.86±0.39	22.76±0.45
LG-FedAvg [23]	72.83±0.28	70.44±0.31	67.55±0.09	33.65±0.19	27.13±0.37	24.82±0.28
FedPer [2]	77.84±0.18	77.76±0.22	75.01±0.20	35.22±0.67	28.63±0.70	25.56±0.26
Per-FedAvg [7]	75.49±0.74	76.27±0.50	75.41±0.35	32.89±0.43	32.24±0.75	32.59±0.21
FedRoD [4]	79.73±0.68	79.61±0.22	77.76±0.32	39.55±0.58	33.87±2.42	31.49±0.19
FedBABU [28]	78.92±0.36	79.35±0.84	76.34±0.22	32.71±0.23	29.66±0.64	27.72±0.11
PGFed (ours)	81.02 ± 0.41	81.42 ± 0.31	78.56 ± 0.35	43.12 ± 0.03	38.45 ± 0.44	35.71 ± 0.54
PGFedMo (ours)	**81.20±0.08**	**81.48±0.32**	**78.74±0.22**	**43.44±0.14**	**38.50±0.45**	**35.76±0.65**

✓ **PGFed and PGFedMo boost the accuracy by up to 15.47%.**

PGFed – Experiments & results

- Convergence speed
  - CIFAR10

	25 clients #round	25 clients speedup	50 clients #round	50 clients speedup	100 clients #round	100 clients speedup
APFL	31	1.0×	28	1.7×	24	2.6×
FedPer	8	3.9×	6	7.8×	8	7.9×
Per-FedAvg	31	1.0×	47	1.0×	63	1.0×
FedRoD	26	1.2×	35	1.3×	10	6.3×
PGFed	9	3.4×	14	3.4×	15	4.2×
PGFedMo	9	3.4×	14	3.4×	15	4.2×

- Mean individual gain over `Local`
  - CIFAR10

	25 clients	50 clients	100 clients
FedAvg	-8.99 ± 10.36	-8.90 ± 15.48	-5.02 ± 14.30
APFL	2.79 ± 8.07	5.73 ± 8.43	8.37 ± 6.91
FedPer	5.31 ± 2.56	8.31 ± 6.00	8.63 ± 5.26
Per-FedAvg	0.72 ± 6.22	5.02 ± 7.39	8.09 ± 7.00
FedRoD	7.80 ± 3.68	8.84 ± 6.29	10.68 ± 6.14
PGFed	8.49 ± 4.67	10.78 ± 5.88	11.15 ± 5.06
PGFedMo	8.61 ± 3.59	10.90 ± 6.11	11.16 ± 5.44

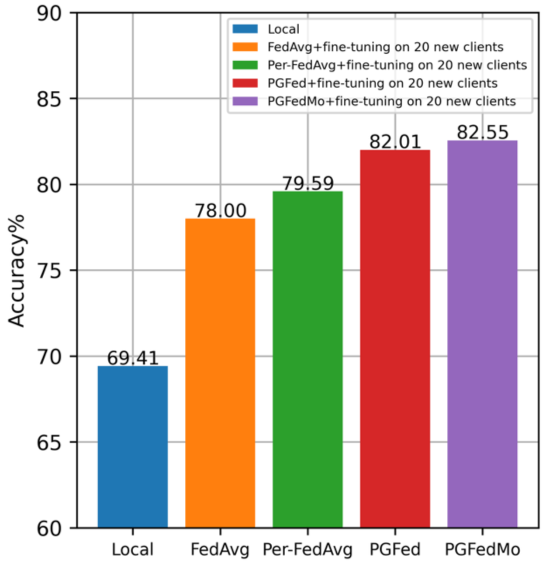
  - CIFAR100

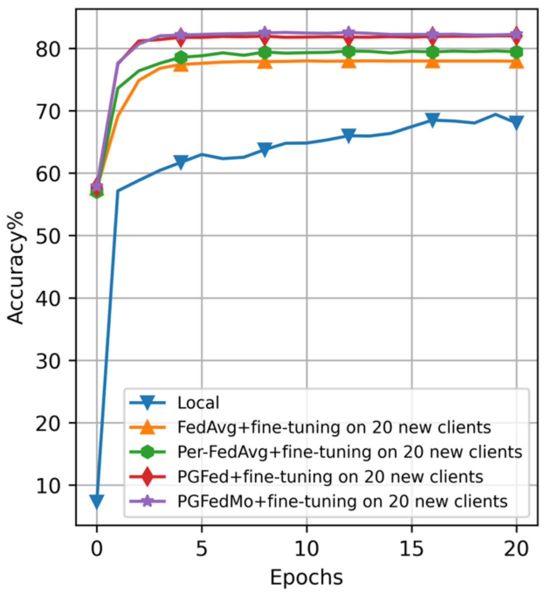
	25 clients	50 clients	100 clients
FedAvg	-3.29 ± 4.22	0.02 ± 4.63	1.77 ± 6.38
APFL	6.48 ± 2.93	8.70 ± 3.37	9.31 ± 4.55
FedPer	3.43 ± 1.80	2.16 ± 2.45	2.31 ± 3.54
Per-FedAvg	0.07 ± 3.71	5.47 ± 3.86	7.49 ± 5.73
FedRoD	7.32 ± 2.68	6.59 ± 3.17	7.47 ± 3.69
PGFed	9.34 ± 1.71	9.01 ± 2.97	12.05 ± 3.93
PGFedMo	9.40 ± 1.87	8.99 ± 2.76	12.07 ± 3.97

✓ **PGFed and PGFedMo have 3.7× average speedup with highest individual gain.**

PGFed – Experiments & results

- Adaptive ability on new clients
  - CIFAR10 & CIFAR100
  - FL on 80 clients, fine-tune global model for 20 epochs on 20 new clients
  - Mean personalized acc. on 20 new clients

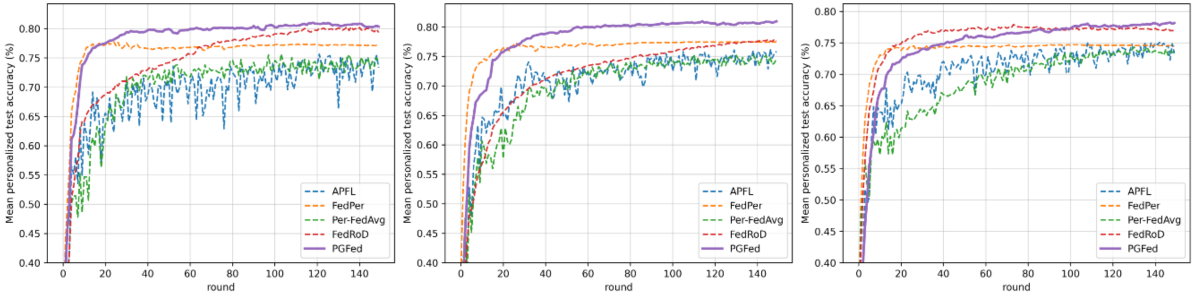




✓ **Global models of PGFed and PGFedMo have highest generailizability**

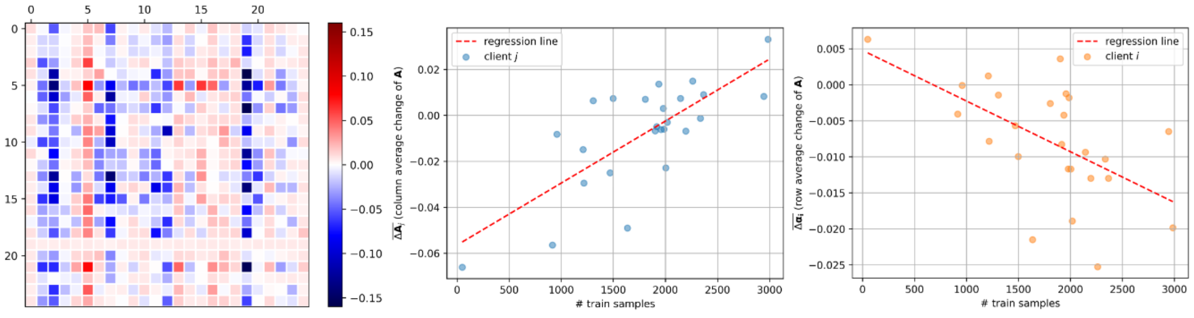
PGFed – More experiments & results

Details in paper



(a) 25 clients (b) 50 clients (c) 100 clients

Figure 1. Convergence behavior of the personalized FL approaches with top performance on CIFAR10. While achieving the highest accuracy performance, PGFed is also able to consistently converge faster than several of the baselines that reach high accuracies.



(a) (b) (c)

Figure 2. Visualization of the change in $\boldsymbol{A}$. Figure (a) is a heat map of the change in $\boldsymbol{A}$. For Figure (b) and (c), the Y-axis of Figure (b) represents the column average of the change in $\boldsymbol{A}$ (the average change of weights of client j's empirical risk on other clients). The Y-axis of Figure (c) is the row average of the change in $\boldsymbol{A}$ (the average change of weights of the auxiliary risk on client i). Through the regression line, we verify the positive correlation between $\overline{\Delta \boldsymbol{A}_j}$ and n_j in Figure (b), and the negative correlation between $\overline{\Delta \boldsymbol{\alpha}_i}$ and n_i in Figure (c).

	Art	Clipart	Product	Real World	Mean
Local	17.16 ± 0.85	37.65 ± 0.47	43.83 ± 0.40	24.50 ± 0.21	30.79 ± 0.23
FedAvg	11.68 ± 1.26	41.29 ± 0.85	42.49 ± 1.28	19.14 ± 0.89	28.65 ± 0.49
APFL	19.11 ± 1.55	44.67 ± 0.61	**50.40 ± 0.56**	25.85 ± 0.88	35.00 ± 0.41
FedRep	20.24 ± 1.45	38.43 ± 1.02	43.70 ± 1.04	24.02 ± 0.81	31.60 ± 0.05
LGFedAvg	17.54 ± 0.45	38.75 ± 0.13	44.59 ± 0.62	25.79 ± 0.61	31.67 ± 0.21
FedPer	17.83 ± 1.07	38.97 ± 0.35	45.87 ± 0.13	25.01 ± 0.52	31.92 ± 0.24
Per-FedAvg	14.62 ± 0.40	39.94 ± 1.29	44.40 ± 1.32	21.58 ± 0.65	30.13 ± 0.07
FedRoD	19.67 ± 1.23	42.44 ± 0.77	44.34 ± 2.07	24.28 ± 1.69	32.68 ± 0.69
FedBABU	18.18 ± 3.54	42.10 ± 2.31	43.51 ± 0.91	**26.81 ± 1.86**	33.38 ± 0.29
PGFed	**22.40 ± 0.26**	**46.48 ± 1.00**	49.86 ± 2.14	26.04 ± 0.80	**36.19 ± 0.92**
PGFedMo	22.16 ± 0.45	45.88 ± 0.83	49.45 ± 0.19	26.60 ± 0.99	36.02 ± 0.20

Table 2. Mean and standard deviation over three trials of the mean personalized accuracy% of the four domains (5 clients/domain) and the average performance on Office-home dataset. The highest and second-highest accuracies under each setting are in **bold** and underlined, respectively.

	25 clients sample 50% Dir(1.0)	50 clients sample 25% Dir(0.3)	100 clients sample 25% Dir(0.3)
Local	90.45±0.19	90.63±0.07	87.14±0.10
FedAvg	99.11±0.03	98.74±0.04	98.47±0.08
APFL	97.49±0.05	97.53±0.06	96.19±0.11
FedRep	95.06±0.16	94.86±0.07	92.47±0.04
LGFedAvg	90.47±0.18	90.99±0.08	87.52±0.22
FedPer	97.89±0.06	97.55±0.08	95.56±0.33
Per-FedAvg	98.40±0.02	96.80±0.04	95.09±0.07
FedRoD	98.61±0.05	98.14±0.09	97.05±0.06
FedBABU	96.49±0.28	94.33±0.13	91.07±0.23
PGFed	99.20±0.04	**99.17±0.05**	**98.94±0.02**
PGFedMo	**99.21±0.04**	99.17±0.07	98.86±0.06

Table 1. Mean and standard deviation over three trials of the mean personalized test accuracy (%) on OrganAMNIST

	Images/s	Relative speed	Accuracy
FedAvg	6917.1	100.00%	64.41±0.66
APFL	3389.8	48.99%	77.36±0.18
Per-FedAvg	3464.5	50.09%	76.27±0.50
FedRoD	6682.4	96.61%	79.61±0.22
PGFed	6120.0	88.48%	81.42±0.31
PGFedMo	6032.8	87.22%	81.48±0.32
PGFed-CE*	6175.5	89.28%	81.16±0.56

* A more communication-efficient variation of PGFed, introduced in Appendix D

Table 3. Computational speed (in terms of "images/s") and accuracy on CIFAR10 with 50 clients

Overview

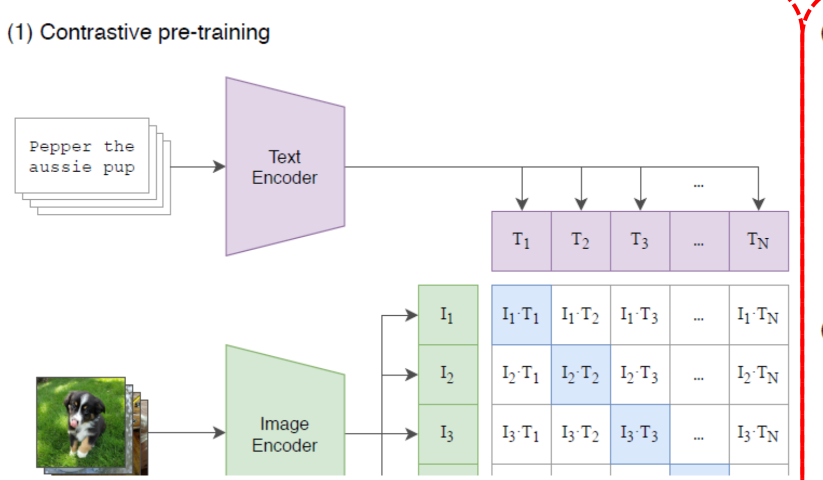
- Federated learning: introduction

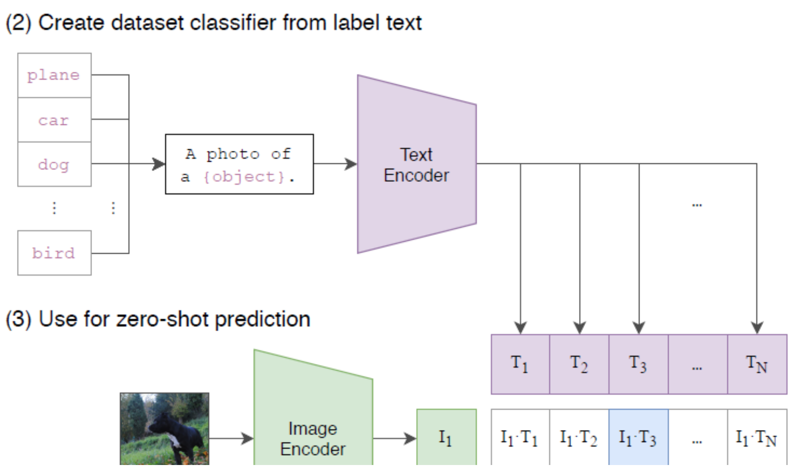
- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- PFL ICLR '25 — Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

- Summary

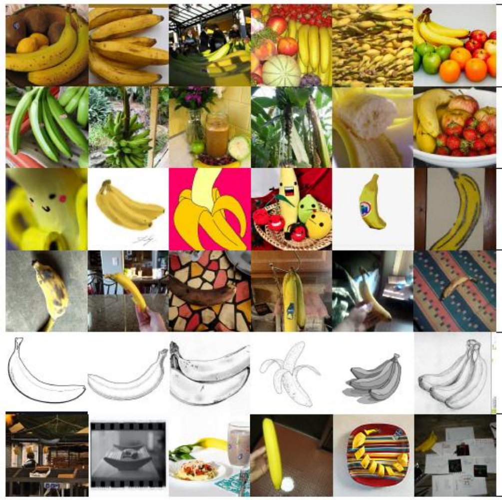
pFedMoAP – Background and motivation

- Vision-Language Models (VLMs) like CLIP with their robust representation learning capabilities, show promise for addressing data heterogeneity in federated learning.





Dataset Examples

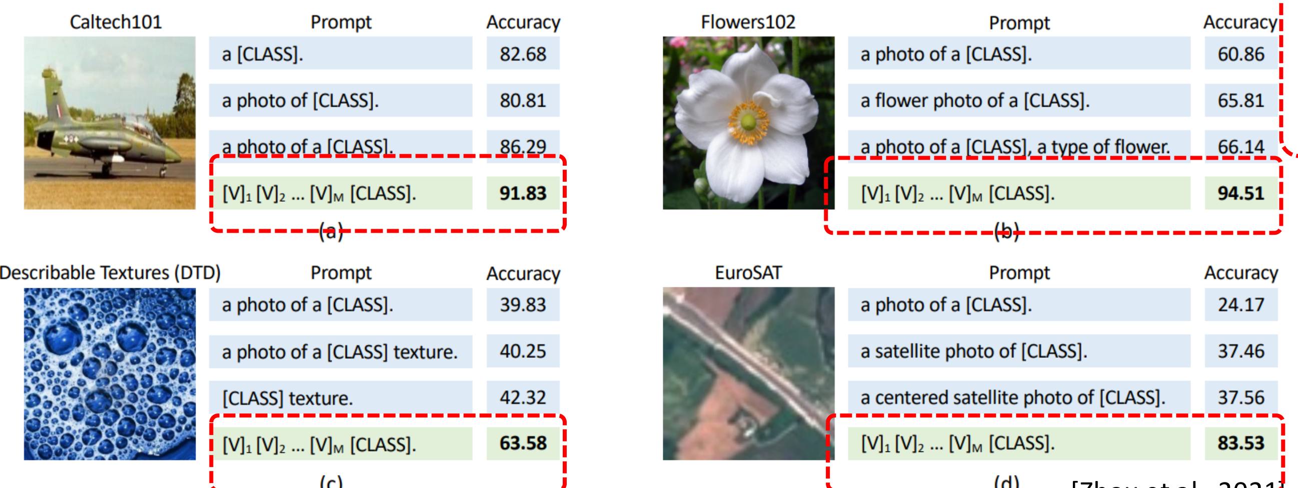


	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%

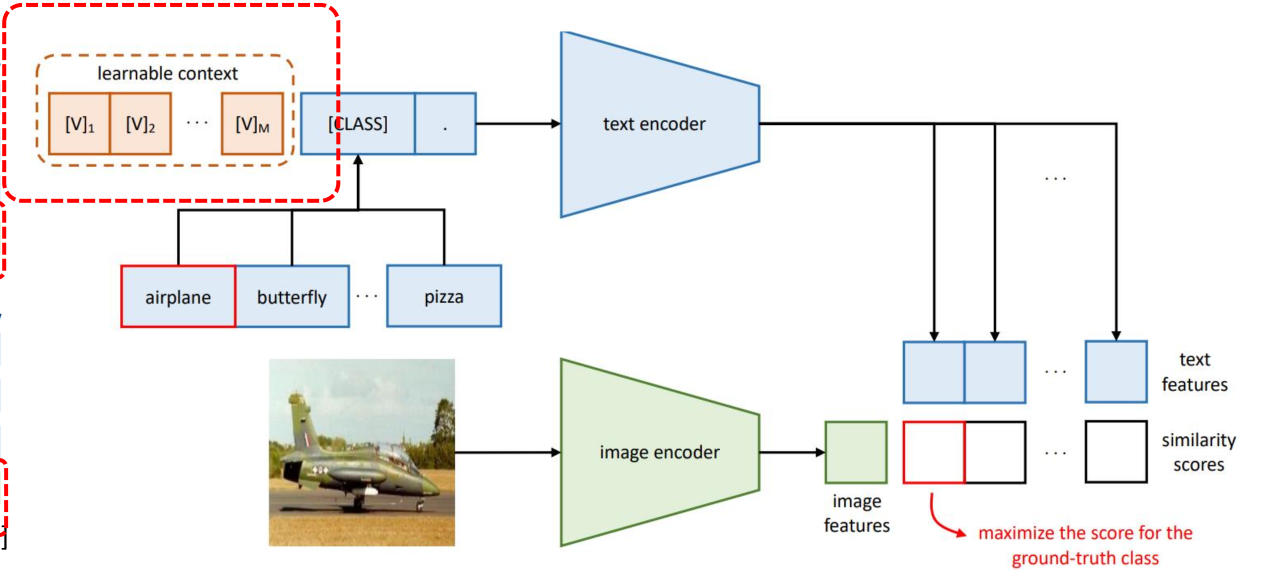
[Figures from CLIP paper]

pFedMoAP – Background and motivation

- Vision-Language Models (VLMs) like CLIP with their robust representation learning capabilities, show promise for addressing data heterogeneity in federated learning.
- Traditional fine-tuning of VLMs in federated settings is challenging due to high communication overhead, leading researchers to explore prompt learning as a more efficient adaptation technique.



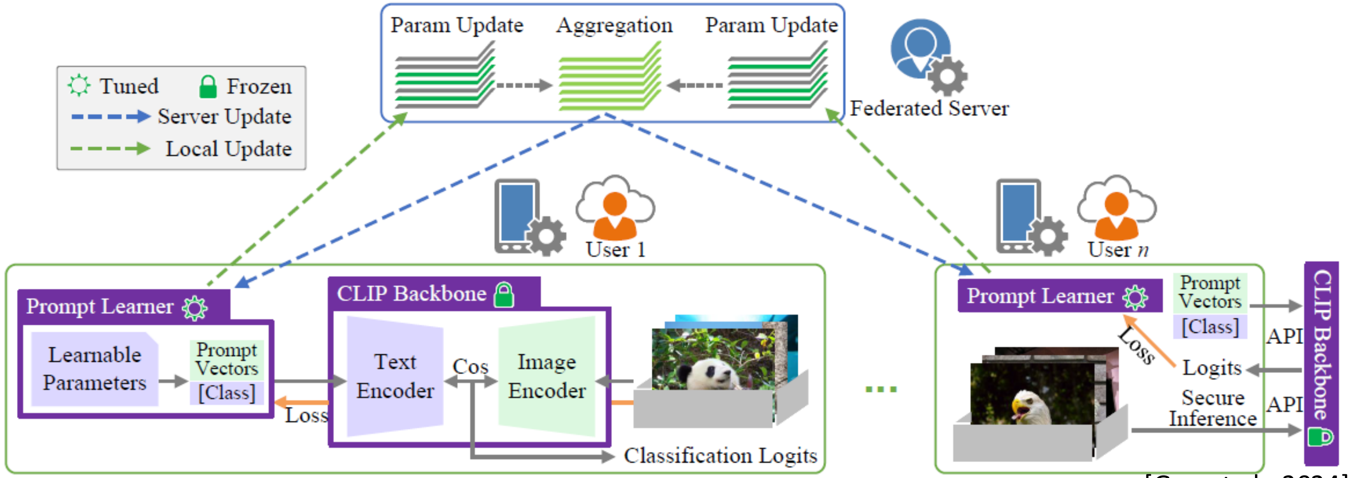
[Zhou et al., 2021]



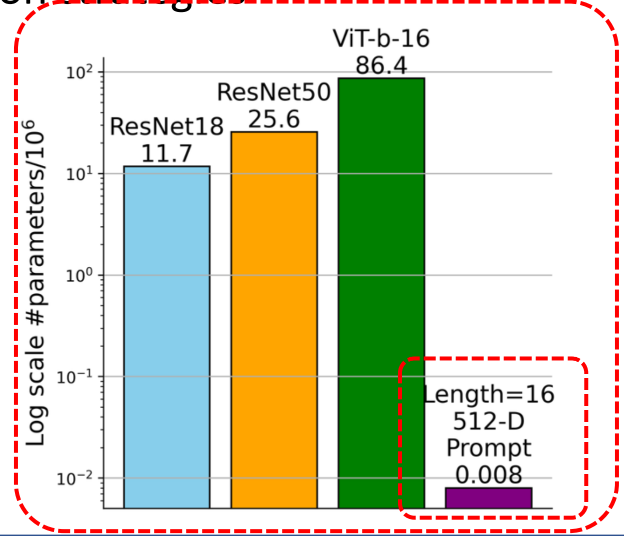
[Zhou et al., 2021]

pFedMoAP – Background and motivation

- Vision-Language Models (VLMs) like CLIP with their robust representation learning capabilities, show promise for addressing data heterogeneity in federated learning.
- Traditional fine-tuning of VLMs in federated settings is challenging due to high communication overhead, leading researchers to explore prompt learning as a more efficient adaptation technique.
- Existing federated prompt learning works
  - Habitually fall into traditional FL paradigm where clients are restricted to downloading only a single globally aggregated model – not fully leveraging the prompt's lightweight nature
  - Struggling to handle extreme data heterogeneity, lacking personalization strategies

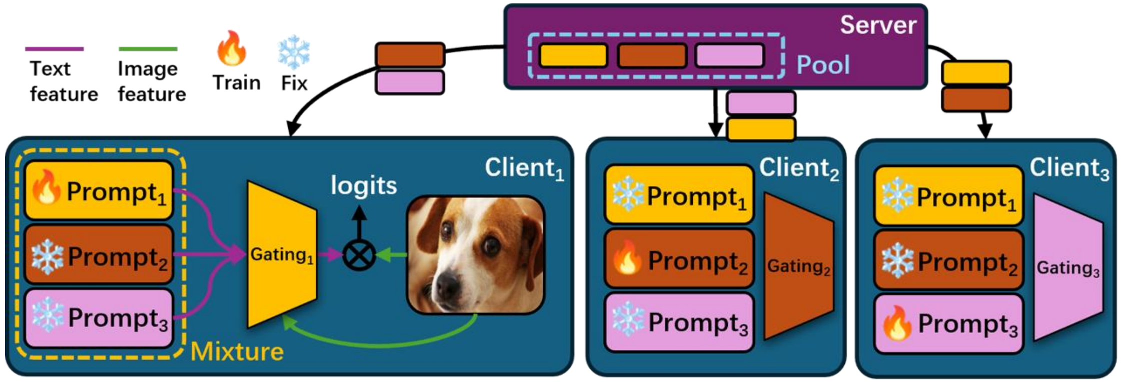


[Guo et al., 2024]



pFedMoAP – Background and motivation

Research question 4: *How can we devise a personalized federated learning framework, tailored for prompt learning in CLIP-like VLMs, while fully exploiting the lightweight nature of the prompts?*



- **P**ersonalized **Fed**erated **M**ixture **o**f **A**daptive **P**rompts (**pFedMoAP**)
  - Allows download of multiple pre-aggregated prompts
  - Uses a Mixture of Experts approach to treat locally updated prompts as specialized experts
  - Implements a client-specific, attention-based gating network to generate enhanced text features

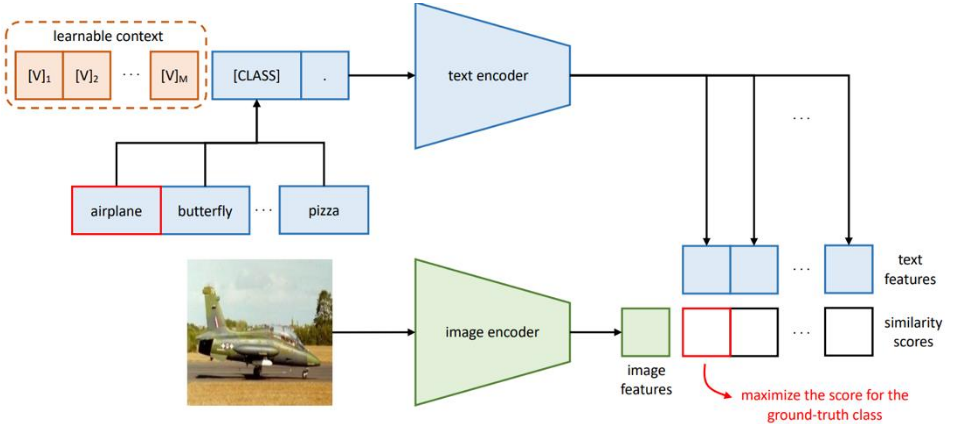
pFedMoAP – Method

- Formulations for existing paradigms
  - Global objective of PFL $\min_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N} F(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N) = \min_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N} \sum_{i=1}^{N} p_i F_i(\boldsymbol{\theta}_i)$
  - Prompt learning for CLIP-like VLMs
    - Learnable prompt $\boldsymbol{P} = \{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_l\} \in \mathbb{R}^{l\times d}$
    - Full prompt $P^{(c)}$ of class c is P with embedding of label c
    - Classification

$$\boxed{\mathrm{logit}^{(c)} = \mathrm{sim}(f(\boldsymbol{x}), g(\boldsymbol{P}^{(c)}))}$$

$$p(\hat{y}=c|\boldsymbol{x}) = \frac{\exp\left(\mathrm{logit}^{(c)}\right)/\tau)}{\sum_{k=1}^{C}\exp\left(\mathrm{logit}^{(k)}/\tau\right)}$$

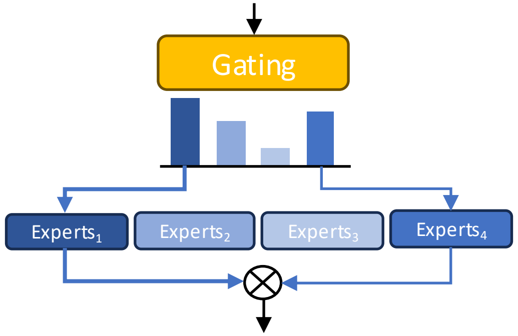
x: image
$f(\cdot)$: CLIP's image encoder
$g(\cdot)$: CLIP's text encoder
τ: temperature



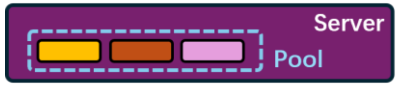
[Figure from CoOp paper]

    - In FL, aggregated global prompt $\boldsymbol{P}_g^t = \sum_{i\in\mathcal{S}_t} \frac{n_i}{\sum_{k\in\mathcal{S}_t} n_k} \boldsymbol{P}_i^t$
  - Mixture of Experts (MoE) output $\boxed{MoE(\boldsymbol{x}) = \sum_{i=1}^{N} G(\boldsymbol{x})_i \cdot E_i(\boldsymbol{x})}$

N: #experts
$E(\cdot)$: an expert
$G(\cdot)$: gating, usually softmax of TopK/N from projected x



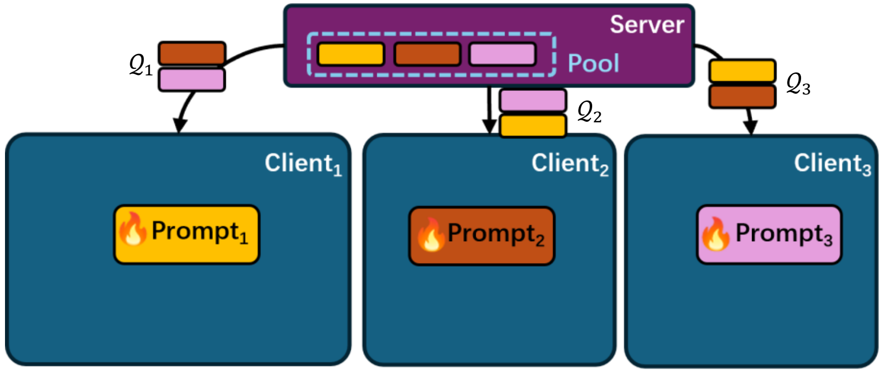
pFedMoAP – Method



- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{\boldsymbol{P}_i^{t-1}\}_{i\in\mathcal{P}_{t-1}\cap\mathcal{S}_t} + \{\boldsymbol{P}_j^{t}\}_{j\in\mathcal{S}_t}$

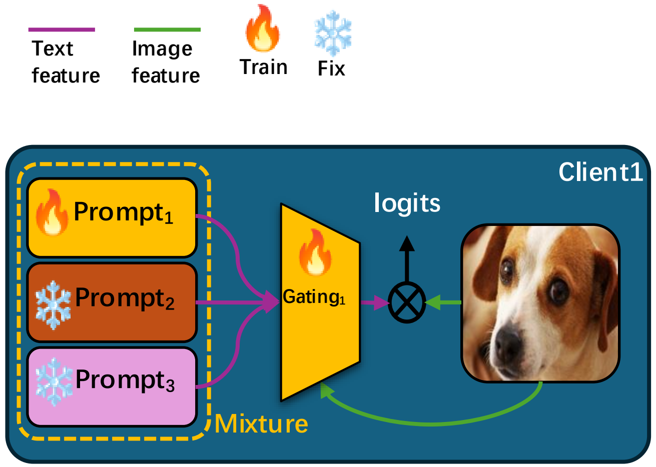
pFedMoAP – Method

- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{P_i^{t-1}\}_{i \in \mathcal{P}_{t-1} \cap \mathcal{S}_t} + \{P_j^t\}_{j \in \mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^K$: set of clients assigned to client i, with prompts P_{NL_j}(NL= abbr. for non-local)



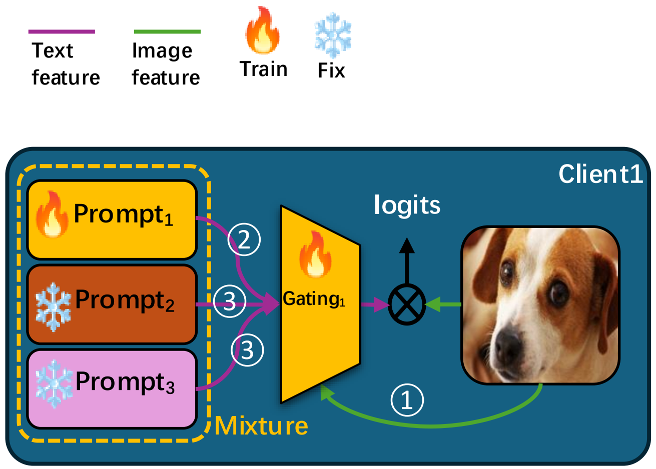
pFedMoAP – Method

- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{P_i^{t-1}\}_{i\in\mathcal{P}_{t-1}\cap\mathcal{S}_t} + \{P_j^t\}_{j\in\mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^{K}$: set of clients assigned to client i, with prompts P_{NL_j}(NL= abbr. for non-local)
  - Before local training, for once, client compute (fixed) text feature from non-local prompts $\forall c \in [C],\ \boldsymbol{T}_{NL}^{(c)} \triangleq \{\boldsymbol{T}_{NL_j}^{(c)} | \boldsymbol{T}_{NL_j}^{(c)} = g(\boldsymbol{P}_{NL_j}^{(c)}),\ \forall NL_j \in \mathcal{Q}_i\}$



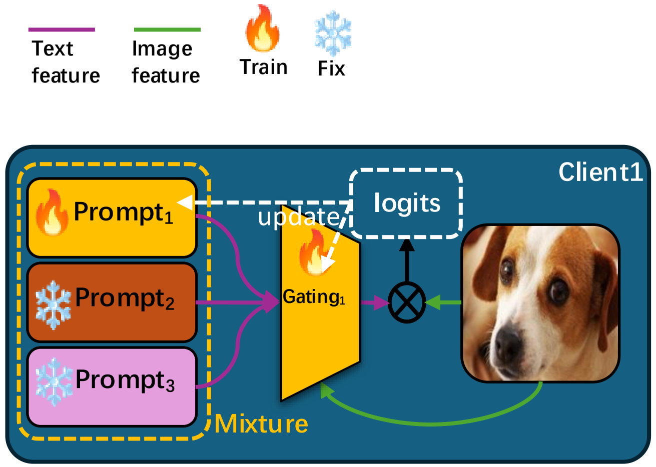
pFedMoAP – Method

- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{P_i^{t-1}\}_{i\in\mathcal{P}_{t-1}\cap\mathcal{S}_t} + \{P_j^{t}\}_{j\in\mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^{K}$: set of clients assigned to client i, with prompts P_{NL_j}(NL= abbr. for non-local)
  - Before local training, for once, client compute (fixed) text feature from non-local prompts $\forall c \in [C],\ T_{NL}^{(c)} \triangleq \{T_{NL_j}^{(c)} | T_{NL_j}^{(c)} = g(P_{NL_j}^{(c)}),\ \forall NL_j \in \mathcal{Q}_i\}$
  - Gating (detailed in following slides)
    - Input type ①: image feature $\boldsymbol{I}_k = f(\boldsymbol{x}_k)$
    - Input type ②: text feature from local prompt $\boldsymbol{T}_L^{(c)} = g(\boldsymbol{P}_i^{(c)})$
    - Input type ③: text features from non-local prompts $\boldsymbol{T}_{NL}^{(c)}$
    - Output: MoE text feature $\forall c \in [C],\ \boldsymbol{T}_{MoE}^{(c)} \triangleq G(\boldsymbol{I}_k, \boldsymbol{T}_L^{(c)}, \boldsymbol{T}_{NL}^{(c)} | \boldsymbol{\theta}_i)$



pFedMoAP – Method

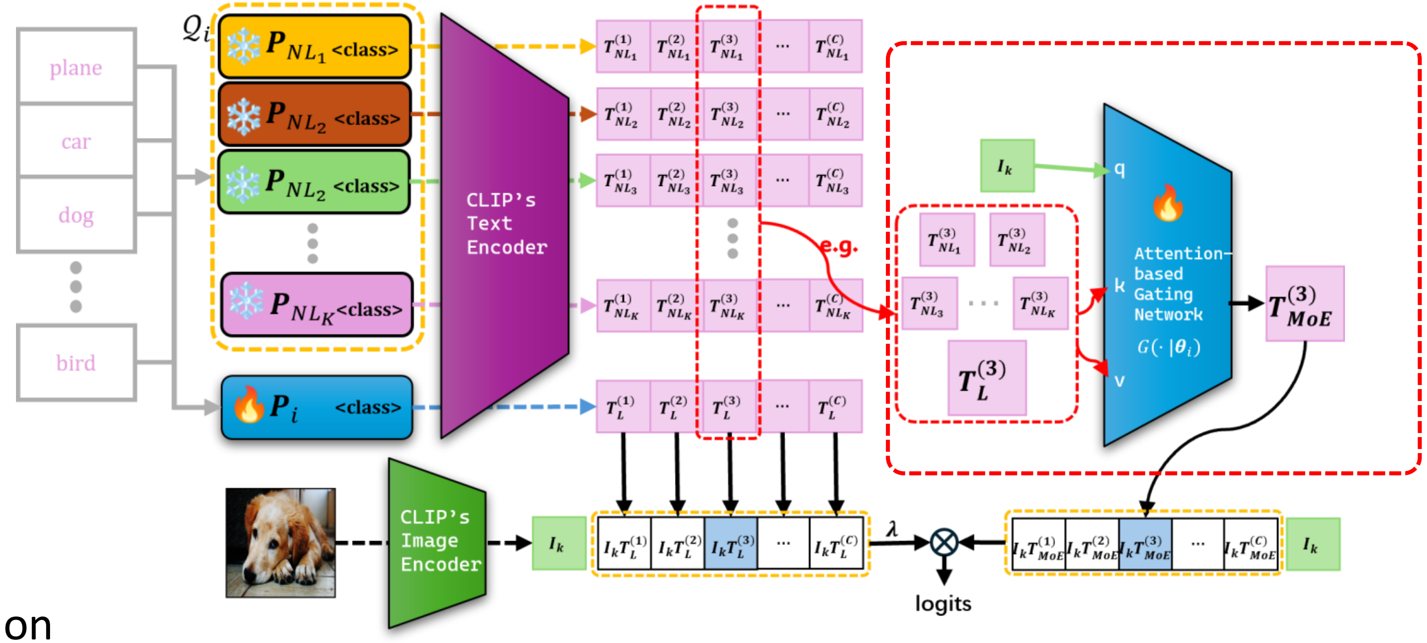
- Workflow
  - Server maintains a pool of prompts $\mathcal{P}_t = \mathcal{P}_{t-1} - \{P_i^{t-1}\}_{i\in\mathcal{P}_{t-1}\cap\mathcal{S}_t} + \{P_j^t\}_{j\in\mathcal{S}_t}$
  - Each client $i \in S_t$ download K pre-aggregated (non-local) prompt
    - K-Nearest Neighbors (KNN) since most likely to have similar distribution
    - $\mathcal{Q}_i = \{NL_j\}_{j=1}^{K}$: set of clients assigned to client i, with prompts P_{NL_j}(NL= abbr. for non-local)
  - Before local training, for once, client compute (fixed) text feature from non-local prompts $\forall c \in [C],\ T_{NL}^{(c)} \triangleq \{T_{NL_j}^{(c)} | T_{NL_j}^{(c)} = g(P_{NL_j}^{(c)}),\ \forall NL_j \in \mathcal{Q}_i\}$
  - Gating (detailed in following slides)
    - Input type 1: image feature $I_k = f(x_k)$
    - Input type 2: text feature from local prompt $T_L^{(c)} = g(P_i^{(c)})$
    - Input type 3: text features from non-local prompts $T_{NL}^{(c)}$
    - Output: MoE text feature $\forall c \in [C],\ T_{MoE}^{(c)} \triangleq G(I_k, T_L^{(c)}, T_{NL}^{(c)} | \theta_i)$
  - Final step: compute logits, manually address local prompt since it is the only locally learnable prompt $\forall c \in [C],\ \mathrm{logit}^{(c)} = \mathrm{sim}(\boldsymbol{I}_k, \boldsymbol{T}_{MoE}^{(c)}) + \lambda \cdot \mathrm{sim}(\boldsymbol{I}_k, \boldsymbol{T}_L^{(c)})$



University of Pittsburgh

pFedMoAP – Method

- Attention-based gating network: mechanism

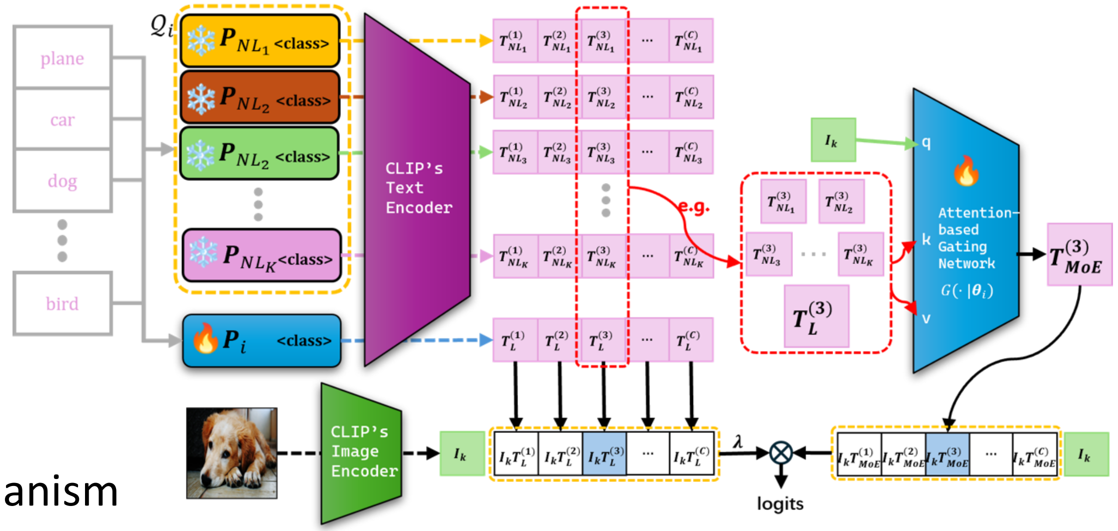


- Multi-head attention
- Pooling on features to reduce the size of gating from 1024 to 128
- Q=Pooling(I_k), K=V=Pooling{ $T_L^{(c)}, T_{NL_1}^{(c)}, T_{NL_2}^{(c)}, \dots, T_{NL_K}^{(c)}$ }
- MoE text feature: $T_{MoE}^{(c)} = G(I_k, T_L^{(c)}, T_{NL}^{(c)}|\theta_i) = \mathrm{MHA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \dots, \mathrm{head}_h)W^O \quad \mathrm{head}_q = \mathrm{Attention}(QW_q^Q, KW_q^K, VW_q^V)$

pFedMoAP – Method

- Attention-based gating network: design rationale against traditional projection-based gating network
  - Projection-based gating network $G_{\text{proj}}(\boldsymbol{x}_k) \in \mathbb{R}^{K+1}$

    $$MoE(\boldsymbol{x}) = \sum_{i=1}^{N} G(\boldsymbol{x})_i \cdot E_i(\boldsymbol{x})$$
  - Attention-based gating against projection-based gating
    - is more robust to adaptive experts
    - serves as linear probing with more capacity
    - leverages CLIP's feature alignment with attention mechanism
    - is agnostic to experts' order



pFedMoAP – Experiments & results

- Datasets

Dataset	Training Set Size	Test Set Size	Number of Classes	Number of Clients	Sample Rate	Data Heterogeneity
Flowers102	4,093	2,463	102	10	100%	Pathological non-IID
OxfordPets	2,944	3,669	37	10	100%	Pathological non-IID
Food101	50,500	30,300	101	10	100%	Pathological non-IID
Caltech101	4,128	2,465	100	10	100%	Pathological non-IID
DTD	2,820	1,692	47	10	100%	Pathological non-IID
Office-Caltech10	2,025	508	10	20	50%	Dir(0.3)
DomainNet	18,278	4,573	10	30	25%	Dir(0.3)
CIFAR10	50,000	10,000	10	100	10%	Dir(0.5)
CIFAR100	50,000	10,000	100	100	10%	Dir(0.5)

CLIP datasets, pathological label shift

Domain adaptation datasets, feature + label shift

CIFAR 10/100, Practical label shift

pFedMoAP – Experiments & results

- Datasets

Dataset	Training Set Size	Test Set Size	Number of Classes	Number of Clients	Sample Rate	Data Heterogeneity
Flowers102	4,093	2,463	102	10	100%	Pathological non-IID
OxfordPets	2,944	3,669	37	10	100%	Pathological non-IID
Food101	50,500	30,300	101	10	100%	Pathological non-IID
Caltech101	4,128	2,465	100	10	100%	Pathological non-IID
DTD	2,820	1,692	47	10	100%	Pathological non-IID
Office-Caltech10	2,025	508	10	20	50%	Dir(0.3)
DomainNet	18,278	4,573	10	30	25%	Dir(0.3)
CIFAR10	50,000	10,000	10	100	10%	Dir(0.5)
CIFAR100	50,000	10,000	100	100	10%	Dir(0.5)

- Compared methods
  - Local methods
    - Zero-shot CLIP
    - CoOp (prompt learning)
  - Federated prompt learning + FL/PFL
    - PromptFL
    - PromptFL + FedProx
    - PromptFL + FT
    - PromptFL + FedAMP
    - PromptFL + FedPer
  - Personalization designed for federated prompt learning
    - pFedPrompt
    - FedOTP

pFedMoAP – Experiments & results

- Pathological label shift on CLIP datasets

	Flowers102	OxfordPets	Food101	Caltech101	DTD
ZS-CLIP [71]	62.17±0.12	84.47±0.01	75.27±0.05	85.14±0.24	40.21±0.12
CoOp [100]	70.14±0.76	83.21±1.30	70.43±2.42	87.37±0.44	44.23±0.63
PromptFL [31]	72.80±1.14	90.79±0.61	77.31±1.64	89.70±1.99	54.11±0.22
PromptFL+FT [12]	72.31±0.91	91.23±0.50	77.16±1.56	89.70±0.25	53.74±1.36
PromptFL+FedPer [5]	72.11±1.35	89.50±1.62	71.29±1.87	86.72±1.45	50.23±0.82
PromptFL+FedProx [50]	66.40±0.29	89.24±0.41	76.24±1.94	89.41±0.55	44.26±1.11
PromptFL+FedAMP [37]	69.10±0.13	80.21±0.44	74.48±1.71	87.31±1.60	47.16±0.92
pFedPrompt [30]	86.46±0.15	91.84±0.41	92.26±1.34	96.54±1.31	77.14±0.09
FedOTP [48]	96.23±0.44	98.82±0.11	92.73±0.15	97.02±0.36	87.64±0.70
pFedMoAP (λ=0.0)	97.61±0.11	94.83±0.65	86.71±0.15	95.71±0.37	85.64±0.34
pFedMoAP (λ=0.5)	**98.41±0.04**	**99.06±0.09**	**93.39±0.09**	**97.95±0.07**	**89.13±0.54**

pFedMoAP – Experiments & results

- Practical label shift on CIFAR datasets
  - Dir($\alpha = 0.5$), 100 clients, 10% sample rate, 120 rounds
  - CLIP backbone: ResNet50

	CIFAR10	CIFAR10
ZS CLIP [79]	53.46±0.21	32.68±0.00
CoOp [115]	80.84±0.39	48.74±0.17
PromptFL [36]	73.29±0.37	45.00±0.62
Prompt+FedProx [57]	73.32±0.34	45.63±0.75
pFedMoAP	83.46±0.53	53.42±0.22

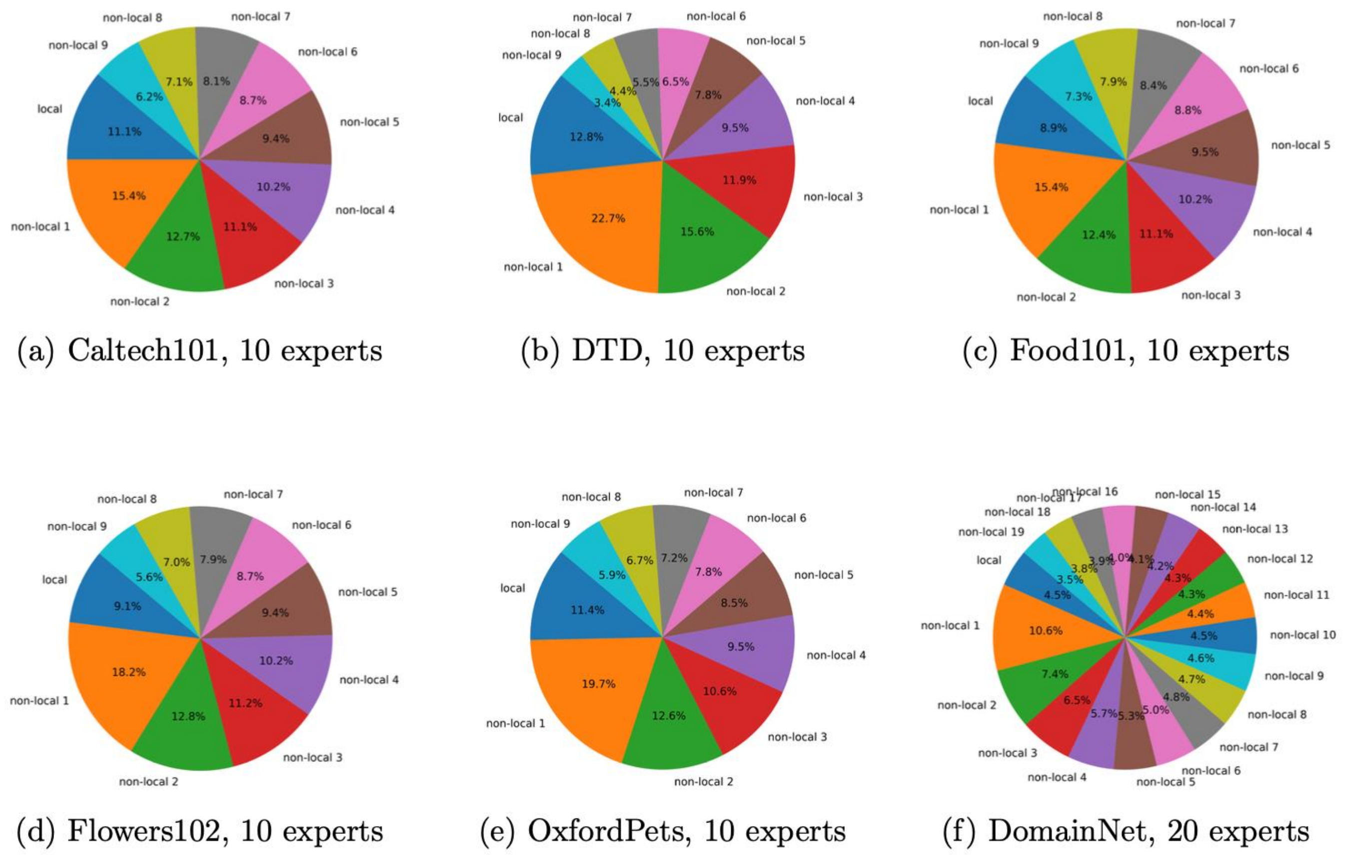
- Feature + label shift on domain adaptation datasets
  - 5 clients/domain, Dir($\alpha = 0.3$)
  - DomainNet = 30 clients, 25% sample rate, 25 rounds
  - Office-Caltech10 = 20 clients, 50% sample rate
  - CLIP backbone: ViT-b-16

DomainNet	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
ZS CLIP	9.18±0.62	10.03±0.16	9.93±0.51	10.25±0.40	9.90±1.30	9.54±1.13	9.81±0.30
CoOp	43.84±3.51	45.72±0.85	29.94±0.46	36.83±1.17	31.64±0.49	33.97±0.78	36.99±0.79
PromptFL	27.63±16.41	27.69±18.07	21.62±8.34	23.45±13.49	20.62±11.03	25.90±8.10	24.48±12.52
Prompt+FedProx	22.23±15.42	21.75±17.00	18.58±8.15	19.40±12.59	17.17±10.25	22.49±8.44	20.27±11.83
pFedMoAP	47.49±0.64	46.73±0.71	32.74±0.84	37.16±0.34	31.02±0.59	37.67±0.72	38.80±0.11

Office-Caltech10	Amazon	Caltech	DSLR	Webcam	Average
ZS-CLIP [79]	9.83±1.63	10.67±0.89	10.89±1.40	6.20±3.84	9.40±0.77
CoOp [115]	30.29±3.64	35.88±1.30	29.89±5.15	33.43±2.25	32.37±1.81
PromptFL [36]	21.08±9.60	23.72±12.21	22.94±7.96	25.88±7.72	23.41±9.06
Prompt+FedProx [57]	18.64±8.58	19.56±11.59	20.89±7.38	22.96±7.56	20.51±8.48
pFedMoAP	35.47±1.37	37.45±1.33	45.11±3.14	35.22±1.04	38.31±1.21

pFedMoAP – Experiments & results

- Contributions of experts



(a) Caltech101, 10 experts
(b) DTD, 10 experts
(c) Food101, 10 experts
(d) Flowers102, 10 experts
(e) OxfordPets, 10 experts
(f) DomainNet, 20 experts

pFedMoAP – Experiments & results

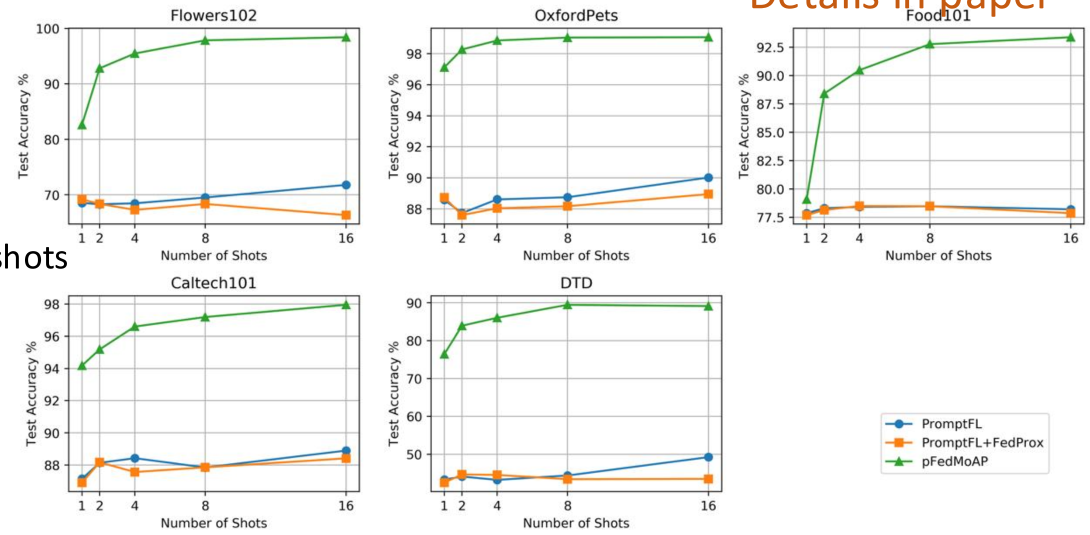
- Attention-based vs. linear projection-base gating network

	Flowers102	OxfordPets	Food101	Caltech101	DTD
Linear projection-based (3 experts)	86.92±1.84	90.54±1.33	78.19±3.07	89.59±1.46	61.42±5.43
Linear projection-based (10 experts)	69.64±4.57	52.78±6.88	77.39±3.29	86.57±1.96	30.42±7.14
Attention-based, with aggregation	97.56±0.07	98.24±0.12	91.89±0.19	96.17±0.18	87.52±0.69
Attention-based, without aggregation (ours)	98.41±0.04	99.06±0.09	93.39±0.09	97.95±0.07	89.13±0.54

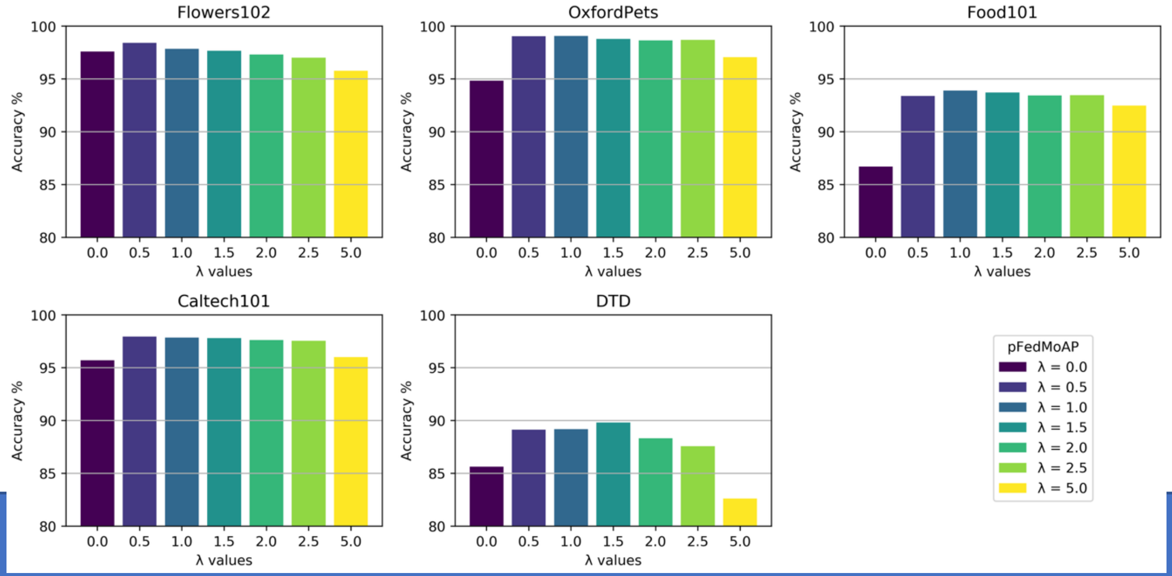
pFedMoAP – Ablation studies

Details in paper

#shots



λ



Differential privacy

	Flowers102	OxfordPets	Food101	Caltech101	DTD
Without differential privacy (from Tab. 11)					
PromptFL [36]	72.80±1.14	90.79±0.61	77.31±1.64	89.70±1.99	54.11±0.22
PromptFL+FedProx [57]	66.40±0.29	89.24±0.41	76.24±1.94	89.41±0.55	44.26±1.11
pFedMoAP (ours)	98.41±0.04	99.06±0.09	93.39±0.09	97.95±0.07	89.13±0.54
With differential privacy ($\epsilon = 50$)					
PromptFL [36]	67.07±0.60	88.05±0.32	77.41±0.60	84.83±0.42	38.39±1.25
PromptFL+FedProx [57]	66.22±0.63	87.78±0.61	77.27±0.59	84.68±0.64	39.43±1.11
pFedMoAP (ours)	98.34±0.06	99.08±0.02	93.36±0.04	97.90±0.08	89.99±0.49
With differential privacy ($\epsilon = 25$)					
PromptFL [36]	64.25±1.10	86.26±1.07	76.84±0.66	85.00±1.59	38.19±0.66
PromptFL+FedProx [57]	62.87±0.99	86.82±0.47	76.21±0.64	84.51±1.52	37.82±0.52
pFedMoAP (ours)	98.36±0.12	99.02±0.04	93.41±0.13	97.99±0.06	89.11±0.28

Feature dimension

	Gating network size	Flowers102	OxfordPets	Food101	Caltech101	DTD
$d_{\text{feature}} = 32$	4.2K	97.28±0.18	98.75±0.32	93.42±0.08	97.37±0.08	88.61±0.89
$d_{\text{feature}} = 64$	16.6K	98.55±0.10	98.91±0.23	93.89±0.12	97.75±0.12	89.96±0.09
$d_{\text{feature}} = 128$	66.0K	98.41±0.04	99.06±0.09	93.39±0.09	97.95±0.07	89.13±0.54
$d_{\text{feature}} = 256$	263.2K	99.01±0.05	98.88±0.21	92.49±0.20	97.93±0.07	90.88±0.16
$d_{\text{feature}} = 512$	1.1M	98.18±0.38	96.85±0.22	90.34±0.31	96.99±0.11	89.65±0.10
$d_{\text{feature}} = 1024$	4.2M	98.11±0.33	95.81±0.84	89.20±0.37	96.82±0.26	89.03±0.14

Overview

- Federated learning: introduction
- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- RSNA 2025 Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection
- Summary

Case study: PFL for real-world breast cancer detection

- Challenges in FL for Medical Imaging
  - Limited sample size
    - Due to high costs of medical imaging and labeling
    - Leads to more severely inconsistent local objectives

Case study: PFL for real-world breast cancer detection

- Challenges in FL for Medical Imaging
  - Limited sample size
  - Data distribution bias
    - Local demographics can hardly represent large population
    - Global distribution often presents extreme imbalance

Case study: PFL for real-world breast cancer detection

- Challenges in FL for Medical Imaging
  - Limited sample size
  - Data distribution bias
  - Both feature and label shift
    - Decease prevalence Geographical regions and demographics
    - Institutional specialization
    - Equipment variations
    - Clinical protocol differences
    - Healthcare access across demographics

Case study: PFL for real-world breast cancer detection

- Challenges in FL for Medical Imaging
  - Limited sample size
  - Data distribution bias
  - Both feature and label shift
  - Uncertainty in labels
    - Subjective nature of medical image interpretation causes challenges in terms of label quality and consistency

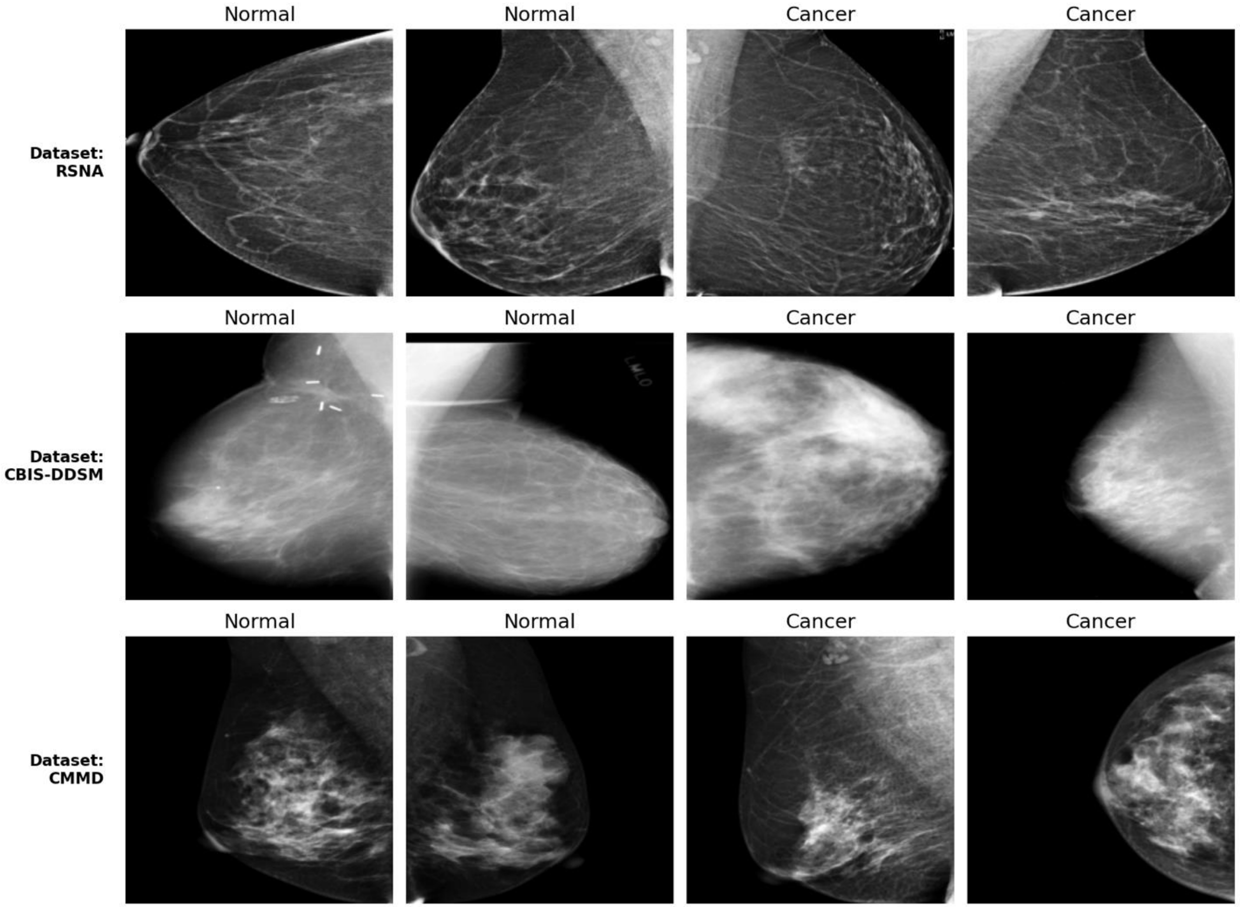
Case study: PFL for real-world breast cancer detection

- Challenges in FL for Medical Imaging
  - Limited sample size
  - Data distribution bias
  - Both feature and label shift
  - Uncertainty in labels

Research question 5: *How can we carefully implement, train, and evaluate existing FL/PFL algorithms and potentially design novel federated frameworks to address the challenges in medical imaging applications?*

Case study: PFL for real-world breast cancer detection

- Datasets
  - RSNA breast cancer detectinon: 54K mammograms with metadata
  - CBIS-DDSM: 3K annotated mammograms with pixel-level lesion masks.
  - CMMD: 5K studies from two Chinese hospitals for cross-domain evaluation.



Case study: PFL for real-world breast cancer detection

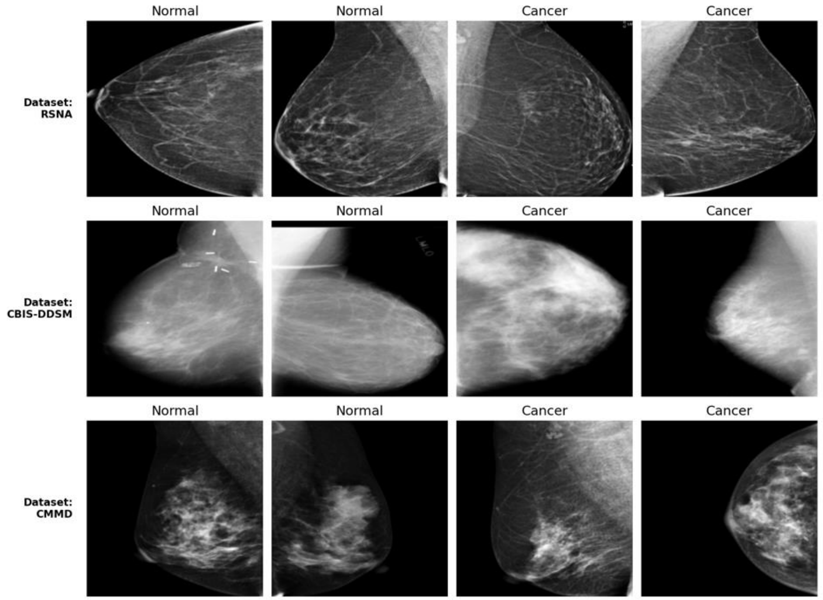
- Data partition for FL
  - **Setting 1:**
    - RSNA only
    - Partitioned by machine ID
  - **Setting 2:**
    - RSNA, CBIS-DDSM, CMMD
    - Partitioned by datasets
  - 85% training, 15% validation

Setting	Client	Dataset Source	Cancer Images	Normal Images
Setting 1	Client 1	RSNA Machine 49	628	2,512
	Client 2	RSNA Machine 48	187	1,122
	Client 3	RSNA Machine 29	184	1,104
	Client 4	RSNA Machine 21	159	1,272
Setting 2	Client 1	RSNA Site A	664	2,656
	Client 2	RSNA Site B	494	2,964
	Client 3	CBIS-DDSM	1,350	1,753
	Client 4	CMMD	4,094	1,108

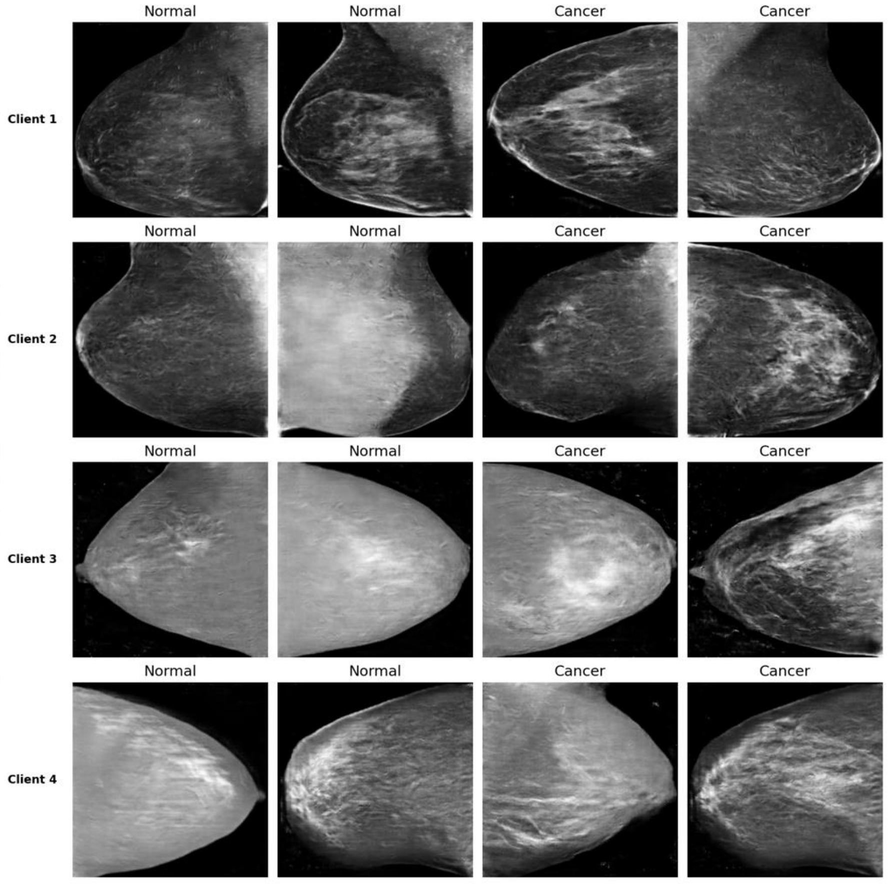
Case study: PFL for real-world breast cancer detection

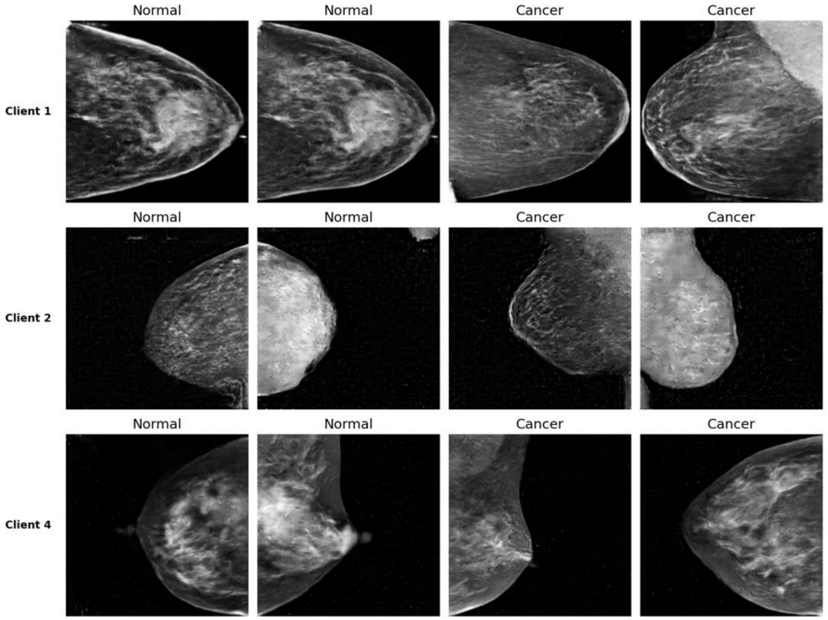
- Methods
  - Local
  - FedAvg
  - APPLE
  - PGFed
- Generative data augmentation
  - Mitigating imbalance
  - Mitigating heterogeneity

Synthetic



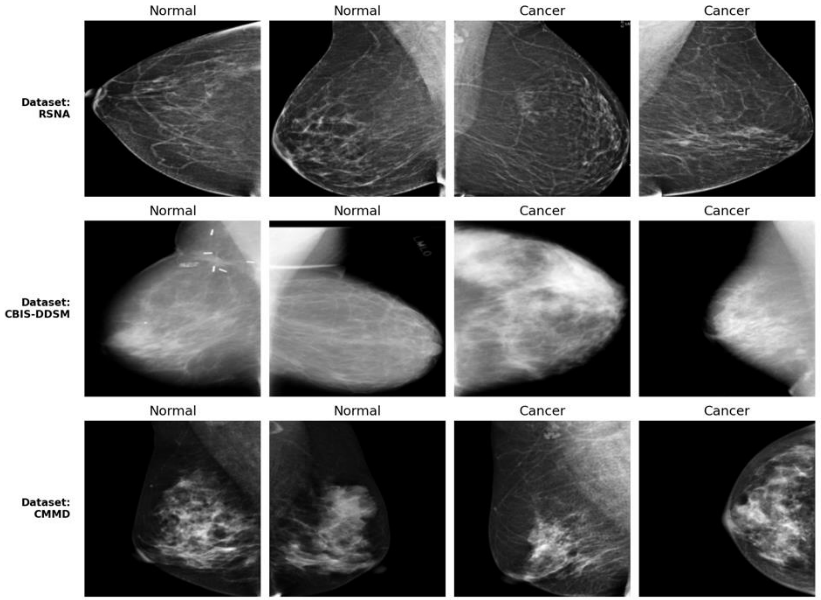
Real



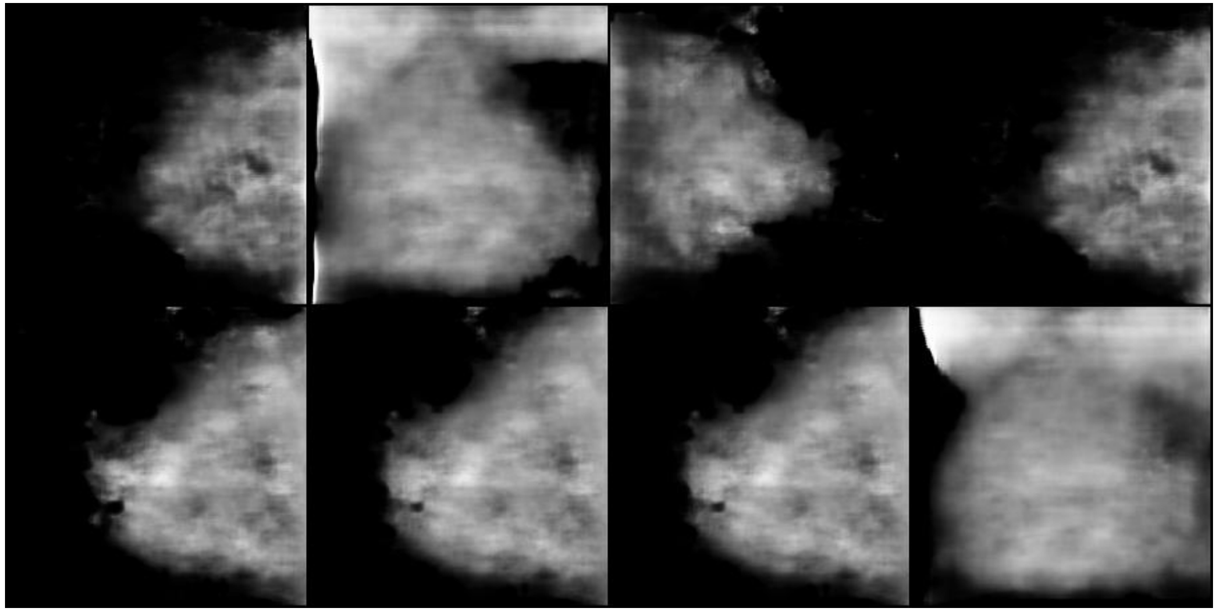


Case study: PFL for real-world breast cancer detection

- Methods
  - Local
  - FedAvg
  - APPLE
  - PGFed
- Generative data augmentation
  - Mitigating imbalance
  - Mitigating heterogeneity



Real



Failed cases: Setting 2 client 2 synthesized images

Case study: PFL for real-world breast cancer detection

- Results

Algorithm	Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
Real data only							
Local	66.59	82.29	64.85 (0.00)	73.37 (0.00)	60.93 (0.00)	67.21 (0.00)	0.00 ± 0.00
FedAvg	63.09	84.26	74.24 (9.39)	60.90 (-12.47)	57.29 (-3.64)	66.99 (-0.22)	-1.73 ± 7.82
APPLE	69.01	85.24	71.92 (7.07)	70.94 (-2.43)	64.48 (3.55)	72.32 (5.10)	3.32 ± 3.55
PGFed	69.32	84.76	74.54 (9.69)	72.25 (-1.12)	66.72 (5.79)	73.95 (6.74)	5.27 ± 3.96

Setting 1: RSNA only

Algorithm	Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
)							
Local	62.37	73.86	60.96 (0.00)	62.47 (0.00)	64.74 (0.00)	61.32 (0.00)	0.00 ± 0.00
FedAvg	64.48	75.43	69.55 (8.59)	65.50 (3.04)	64.49 (-0.25)	62.13 (0.81)	3.05 ± 3.41
APPLE	65.89	76.13	71.51 (10.54)	64.22 (1.75)	69.26 (4.51)	63.28 (1.96)	4.69 ± 3.55
PGFed	66.40	77.13	70.21 (9.24)	69.16 (6.69)	68.33 (3.58)	64.98 (3.66)	5.79 ± 2.35

Setting 2: Multiple datasets

Case study: PFL for real-world breast cancer detection

- Results

Algorithm	Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
Real data only							
`Local`	66.59	82.29	64.85 (0.00)	73.37 (0.00)	60.93 (0.00)	67.21 (0.00)	0.00 ± 0.00
`FedAvg`	63.09	84.26	74.24 (9.39)	60.90 (-12.47)	57.29 (-3.64)	66.99 (-0.22)	-1.73 ± 7.82
`APPLE`	69.01	85.24	71.92 (7.07)	70.94 (-2.43)	64.48 (3.55)	72.32 (5.10)	3.32 ± 3.55
`PGFed`	69.32	84.76	74.54 (9.69)	72.25 (-1.12)	66.72 (5.79)	73.95 (6.74)	5.27 ± 3.96

Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
)						
62.37	73.86	60.96 (0.00)	62.47 (0.00)	64.74 (0.00)	61.32 (0.00)	0.00 ± 0.00
64.48	75.43	69.55 (8.59)	65.50 (3.04)	64.49 (-0.25)	62.13 (0.81)	3.05 ± 3.41
65.89	76.13	71.51 (10.54)	64.22 (1.75)	69.26 (4.51)	63.28 (1.96)	4.69 ± 3.55
66.40	77.13	70.21 (9.24)	69.16 (6.69)	68.33 (3.58)	64.98 (3.66)	5.79 ± 2.35

Setting 1: RSNA only

Setting 2: Multiple datasets

Case study: PFL for real-world breast cancer detection

- Results

Algorithm	Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
Real data only							
Local	66.59	82.29	64.85 (0.00)	73.37 (0.00)	60.93 (0.00)	67.21 (0.00)	0.00 ± 0.00
FedAvg	63.09	84.26	74.24 (9.39)	60.90 (-12.47)	57.29 (-3.64)	66.99 (-0.22)	-1.73 ± 7.82
APPLE	69.01	85.24	71.92 (7.07)	70.94 (-2.43)	64.48 (3.55)	72.32 (5.10)	3.32 ± 3.55
PGFed	69.32	84.76	74.54 (9.69)	72.25 (-1.12)	66.72 (5.79)	73.95 (6.74)	5.27 ± 3.96
Adding synthesized data; FID score for client 1: 125.61, client 2: 220.52, client 3: 194.73, client 4: 184.41							

Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
)						
62.37	73.86	60.96 (0.00)	62.47 (0.00)	64.74 (0.00)	61.32 (0.00)	0.00 ± 0.00
64.48	75.43	69.55 (8.59)	65.50 (3.04)	64.49 (-0.25)	62.13 (0.81)	3.05 ± 3.41
65.89	76.13	71.51 (10.54)	64.22 (1.75)	69.26 (4.51)	63.28 (1.96)	4.69 ± 3.55
66.40	77.13	70.21 (9.24)	69.16 (6.69)	68.33 (3.58)	64.98 (3.66)	5.79 ± 2.35
sized data; FID score for client 1: 120.23, client 2: 200.91, client 3: N/A, client 4: 58.86						

Setting 1: RSNA only

Setting 2: Multiple datasets

Case study: PFL for real-world breast cancer detection

- Results

Algorithm	Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
Real data only							
Local	66.59	82.29	64.85 (0.00)	73.37 (0.00)	60.93 (0.00)	67.21 (0.00)	0.00 ± 0.00
FedAvg	63.09	84.26	74.24 (9.39)	60.90 (-12.47)	57.29 (-3.64)	66.99 (-0.22)	-1.73 ± 7.82
APPLE	69.01	85.24	71.92 (7.07)	70.94 (-2.43)	64.48 (3.55)	72.32 (5.10)	3.32 ± 3.55
PGFed	69.32	84.76	74.54 (9.69)	72.25 (-1.12)	66.72 (5.79)	73.95 (6.74)	5.27 ± 3.96
Adding synthesized data; FID score for client 1: 125.61, client 2: 220.52, client 3: 194.73, client 4: 184.41							
Local	67.06	84.73	69.42 (0.00)	75.31 (0.00)	56.49 (0.00)	63.36 (0.00)	0.00 ± 0.00
FedAvg	64.51	84.49	70.47 (1.05)	55.28 (-20.03)	62.69 (6.20)	71.60 (8.24)	-1.14 ± 11.22
APPLE	67.40	84.89	69.01 (-0.41)	63.69 (-11.62)	66.51 (10.02)	68.94 (5.58)	0.89 ± 8.12
PGFed	70.41	84.76	72.61 (3.19)	68.56 (-6.75)	67.77 (11.28)	74.32 (10.96)	4.67 ± 7.35

Setting 1: RSNA only

Mean AUC	Mean Acc.	Client 1 AUC (gain)	Client 2 AUC (gain)	Client 3 AUC (gain)	Client 4 AUC (gain)	Mean gain ±std
)						
62.37	73.86	60.96 (0.00)	62.47 (0.00)	64.74 (0.00)	61.32 (0.00)	0.00 ± 0.00
64.48	75.43	69.55 (8.59)	65.50 (3.04)	64.49 (-0.25)	62.13 (0.81)	3.05 ± 3.41
65.89	76.13	71.51 (10.54)	64.22 (1.75)	69.26 (4.51)	63.28 (1.96)	4.69 ± 3.55
66.40	77.13	70.21 (9.24)	69.16 (6.69)	68.33 (3.58)	64.98 (3.66)	5.79 ± 2.35
sized data; FID score for client 1: 120.23, client 2: 200.91, client 3: N/A, client 4: 58.86						
63.01	72.89	64.89 (0.00)	60.90 (0.00)	64.74 (0.00)	61.52 (0.00)	0.00 ± 0.00
64.20	74.87	68.23 (3.34)	64.80 (3.90)	65.84 (1.10)	63.26 (1.74)	2.52 ± 1.14
66.38	76.74	67.59 (2.70)	63.52 (2.62)	72.72 (7.98)	65.77 (4.25)	4.39 ± 2.17
67.41	77.03	71.16 (6.27)	71.72 (10.82)	70.38 (5.64)	64.42 (2.90)	6.41 ± 2.85

Setting 2: Multiple datasets

Case study: PFL for real-world breast cancer detection

- Summary
    - Traditional FL (e.g. FedAvg) < Local with severe heterogeneity
    - Proposed PFL (APPLE, PGFed) > FedAvg and Local
    - Generated data: higher perceptive quality usually translate to higher performance

    - For a FL system deployed in real-world
        - Can use more advanced model with better pretraining.
        - Practical hyperparameter tuning is hard, alg's with more hyperparameters is harder to deploy
            - Research: same values for all clients, global metrics for selection.
            - Real-world: clients can use different values, local metrics for selection.
            - Example, hyperparameters: a 2 values, b 3 values,research has 2x3=6 combinations, real-world $a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4$: (2x3)^4=1296 combinations

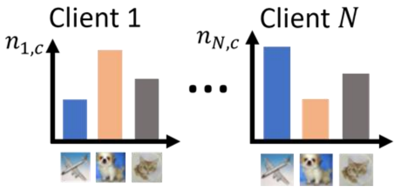
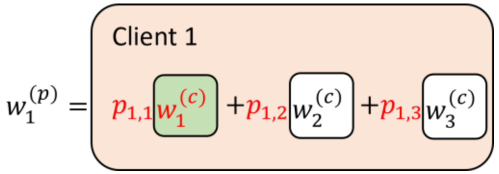
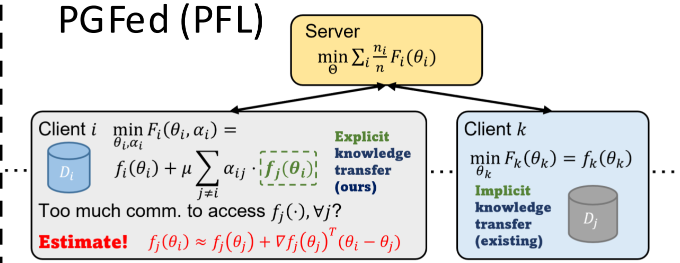
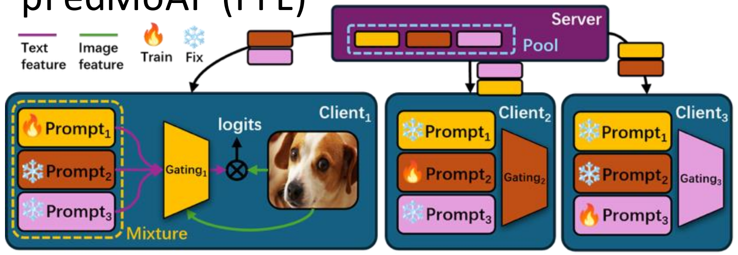
Overview

- Federated learning: introduction

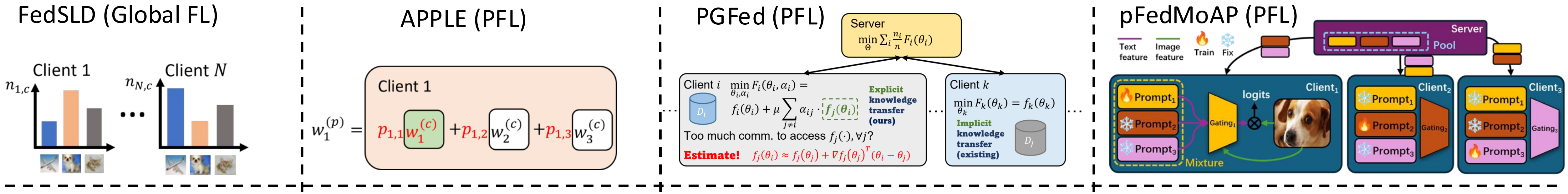
- Federated Learning with Shared Label Distribution for Medical Image Classification (FedSLD)
- Adapt to Adaptation: Learning Personalization for Cross-Silo Federated Learning (APPLE)
- PGFed: Personalize Each Client's Global Objective for Federated Learning (PGFed)
- Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models (pFedMoAP)
- Case Study: Personalized, Real-World, and Cross-Silo Federated Learning for Breast Cancer Detection

- Summary

Summary of the four FL/PFL algorithms

FedSLD (Global FL)	APPLE (PFL)	PGFed (PFL)	pFedMoAP (PFL)
			
Estimate prior and reweighting	PFL with adaptive aggregation	Explicit and efficient personalized global objective with first order approximation	PFL prompt learning for CLIP with attention-based gating network in a MoE structure
+ Sharable label dist. + Reweights sample loss	+ Adaptive personalized aggregation + Flexible between glob./pers. obj's	+ More generalized personalized models + Explicitness with $O(N)$ communication	+ Pre-aggregated prompts sharing allows MoE + Flexible and robust attention-based gating
- Limited performance gain - Only considers label shift	- Large comm. ($O(N^2)$) → cross-silo - Larger memory footprint	- Slightly larger communication than FedAvg - Requires extra server computation/storage	- Clients needs to be able to run CLIP - Computation slightly ↑ as #experts ↑

Summary of the four FL/PFL algorithms



Estimate prior and reweighting

PFL with adaptive aggregation

Explicit and efficient personalized global objective with first order approximation

PFL prompt learning for CLIP with attention-based gating network in a MoE structure

Summary of FL case study: breast cancer detection

- Traditional FL (e.g. FedAvg) < Local with sever heterogeneity
- Proposed PFL (APPLE, PGFed) > FedAvg and Local
- Generated data: higher perceptive quality usually translate to higher performance
- Deploying FL system in real-world will face more challenges (e.g. hyperparameter tuning across heterogeneous clients).

Future directions

- Federated learning with large foundation models.
- Synthetic data generation and augmentation via foundation models.
- Enhancing privacy, security, and trustworthiness in FL.
- Developing multimodal federated systems across diverse domains.

Acknowledgements

- My advisor Dr. Wu and my committee Dr. Zhou, Dr. Jia, and Dr. Tang
- My wife Tianling
- Collaborators: Dr. Chen, Matias and Guangyu from UCF
- ICCI: Oliver, Chang, Zhengbo, Dooman, Jiren, Zhiwei, Giacomo
- Internship: at , and **Sony AI**

Intelligent Computing for Clinical Imaging (ICCI) Lab
-Interfacing computational and medical sciences

Selected funding

- NIH OT (#1OT2OD037972-01)
- NIH/NCI R01 (#CA193603)
- NIH/NCI R01 (#CA218405)
- NSF/NIH joint R01(#EB032896)
- NSF (CICI: SIVD: #2115082)
- NSF (CBET #2229156)
- NIH OT #1OT2OD032701-01
- NIH R01 Supplement (#CA193603-S; #EB032896-03S1)
- UPMC Enterprise (Early Commercialization)
- RSNA Research Scholar Grant (#RSCH1530)
- PA Breast Cancer Coalition
- Jewish Healthcare Foundation
- Pittsburgh Foundation
- Amazon AWS Machine Learning Research Award
- Stanly Marks Research Foundation
- Pitt Momentum Funds Scaling Grant

Thank you!

Jun Luo
jul117@pitt.edu