FedGMKD: An Efficient Prototype Federated Learning Framework through Knowledge Distillation and Discrepancy-Aware Aggregation

Jianqiao Zhang, Caifeng Shan, Jungong Han

Introduction and Motivation

- **Background: Federated Learning (FL) allows collaborative model training while keeping data decentralized, crucial for privacy in fields like healthcare and finance**

- **Challenge: Non-IID data across clients leads to slow convergence and inconsistent performance, making it difficult for global models to generalize**

- **Problem Formulation**

In Federated Learning, each client i has a private dataset D_i and optimizes a local model w_i by minimizing the local loss function:

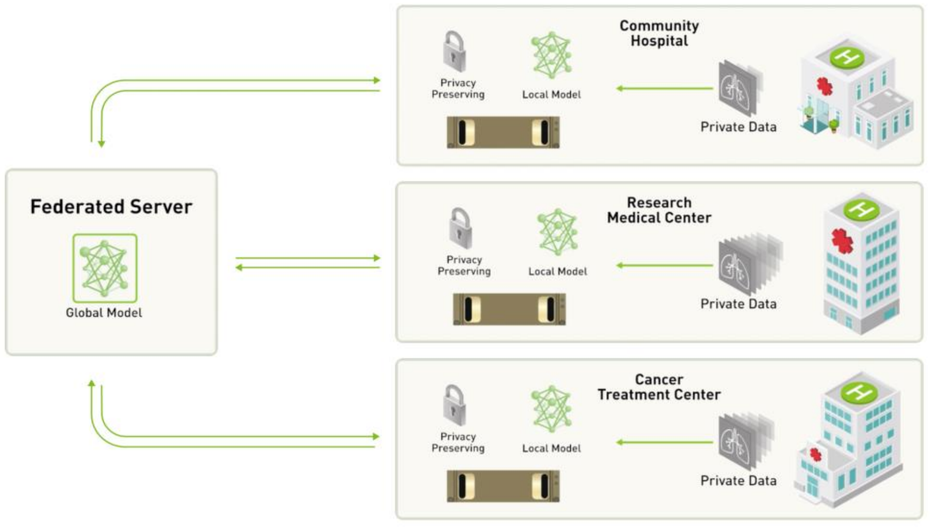
$$F_i(w_i) = \frac{1}{|D_i|} \sum_{x \in D_i} \ell(w_i, x)$$

To build a global model, FedAvg (McMahan et al., 2017) aggregates the local models using weighted averaging:

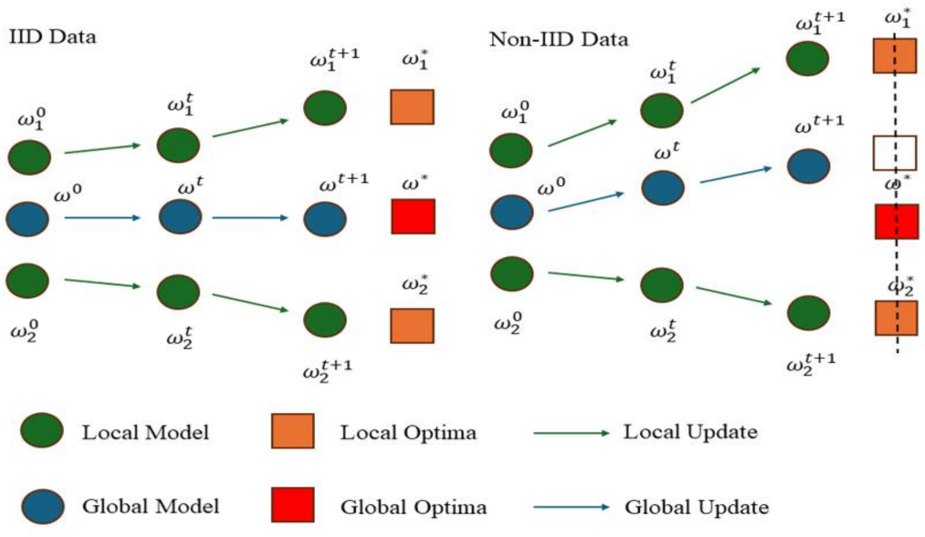
$$W = \frac{1}{N} \sum_{i=1}^{n} |D_i| w_i$$

where $N = \sum_{i=1}^{n} |D_i|$ is the total sample count across all clients.

- **Solution (FedGMKD): We propose FedGMKD, which integrates Cluster Knowledge Fusion (using Gaussian Mixture Models) and Discrepancy-Aware Aggregation. This framework improves model performance by prototype-based distillation knowledge without relying on public datasets and enhances both local and global accuracy across diverse data distributions.**



General Federated Learning Concept (Federal AI, 2023)



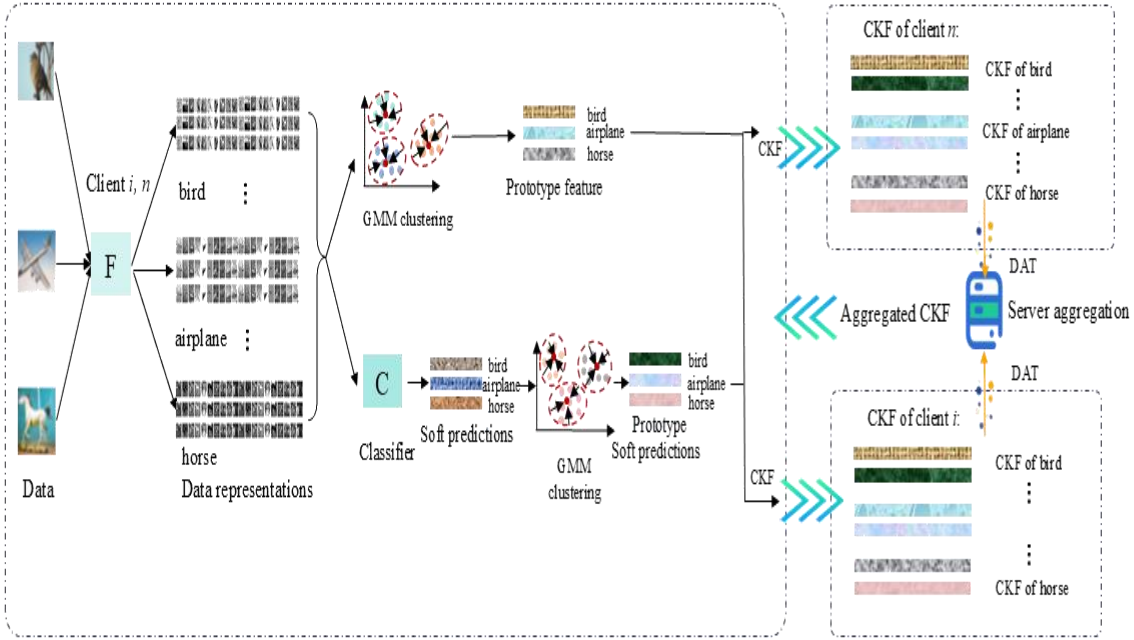
Federated Learning with IID and Non-IID data distribution

- Federal AI. (2023, October 4). Unveiling the dynamics of skin cancer detection with federated learning. Medium. https://medium.com/@federalai/unveiling-the-dynamics-of-skin-cancer-detection-with-federated-learning-f25b590137a7
- McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

Methodology

Cluster Knowledge Fusion (CKF)

- CKF utilizes **Gaussian Mixture Models (GMM)** to generate **prototype features and soft predictions for each class of each client.**
- This method clusters client data to create representative prototypes **without relying on a public dataset**, **preserving data privacy** while **handling non-IID data effectively**



Flow diagram demonstrating the computation of Cluster Knowledge Fusion (CKF) in Federated Learning

- **Feature Extraction and Prediction**

$$h_{L_i} = F_{\theta_i}(x_i),\ z_i = \mathrm{Softmax}\left(C_{\psi_i}(h_i)\right)$$

- **Responsibility Calculation & Prototype Feature and Prediction Calculation**

$$\gamma_m\left(\mathbf{x}_i^j\right) = \frac{\pi_m \cdot \mathcal{N}\left(\mathbf{x}_i^j; \mu_m, \mathbf{\Sigma}_m\right)}{\sum_{s=1}^{M} \pi_s \cdot \mathcal{N}\left(\mathbf{x}_i^j; \mu_s, \mathbf{\Sigma}_s\right)}$$

$$\hat{h}_i^j = \sum_{m=1}^{M} \gamma_m\left(\mathbf{h}_i^j\right)\mu_{m_j},$$

$$\hat{q}_i^f = \sum_{m=1}^{M} \gamma_m\left(\mathbf{z}_i^j\right)\mathbf{z}_{m_j}$$

- **Class-Level CKF for Client**

$$K_i^j = \left(\hat{h}_i^j, \hat{q}_i^j\right)$$

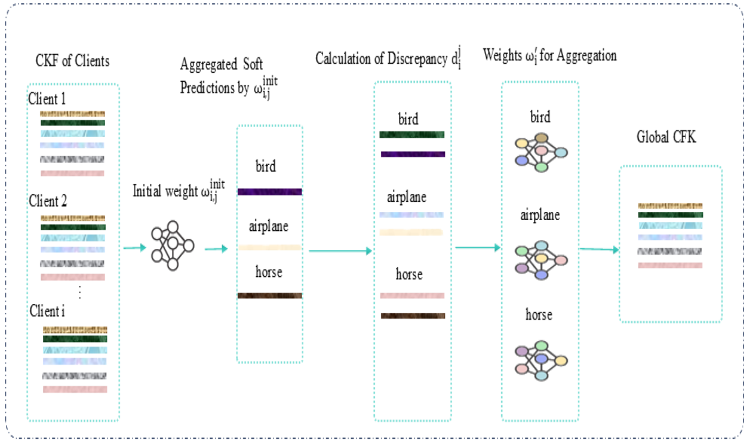
- **Complete CKF for a Client**

$$K_i = \bigcup_{j=1}^{J} \left(\hat{h}_i^j, \hat{q}_i^j\right)$$

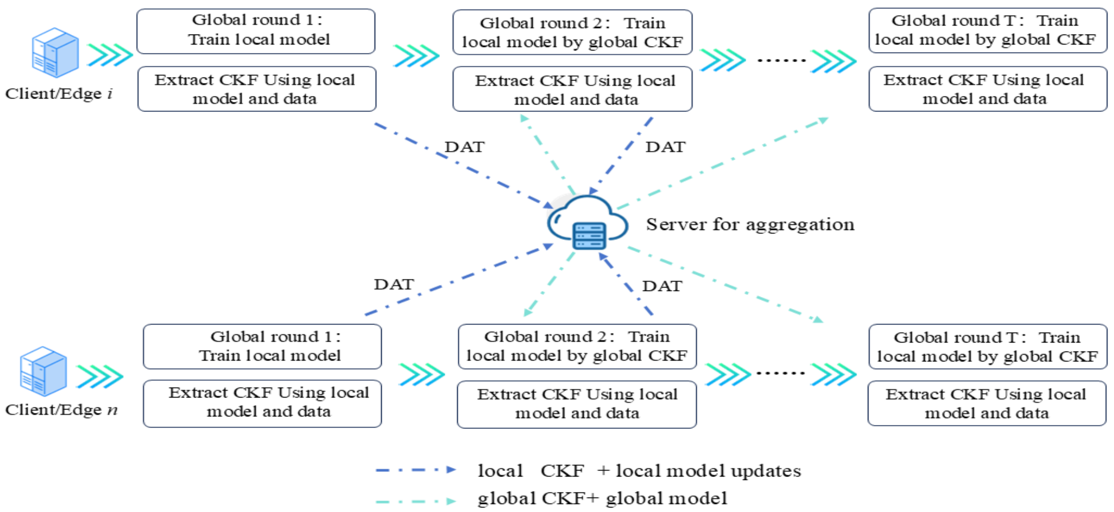
Methodology

Discrepancy-Aware Aggregation Technique (DAT)

- DAT evaluates the **quality** of client data and **weights client contributions** based on **both the quality and quantity of data**



Flow diagram demonstrating the computation of Discrepancy-Aware Aggregation Technique (DAT) in Federated Learning



Flow diagram of overall FedGMKD framework

- **Initial Weight Calculation**

$$w_{i,j}^{\text{init}} = \frac{N_i^j}{\sum_{i=1}^{n} N_i^j}$$

- **KL Divergence for Discrepancy Calculation & Final Weight Adjustment**

$$d_i^j = D_{\text{KL}}\left(\hat{q}_i^j \parallel \hat{Q}_j\right) = \hat{q}_i^j \log \frac{\hat{q}_i^j}{\hat{Q}_j}$$

$$w_i' = \frac{\text{ReLU}\left(w_{i,j}^{\text{init}} - a \cdot d_i^j + b\right)}{\sum_{i=1}^{n} \text{ReLU}\left(w_{i,j}^{\text{init}} - a \cdot d_i^j + b\right)}$$

- **Global CKF Aggregation**

$$\mathbf{H}_j^{r+1} = \sum_{i=1}^{n} w_i' \cdot \hat{h}_i^{j,r}$$

$$\mathbf{Q}_j^{r+1} = \sum_{i=1}^{n} w_i' \cdot \hat{q}_i^{j,r}$$

- **Complete Global CKF**

$$G^{r+1} = \bigcup_{j=1}^{J} \left(\mathbf{H}_j^{r+1}, \mathbf{Q}_j^{r+1}\right)$$

Local Training Objective Function

$$L(\mathcal{D}_i, \mathbf{w}_i) = \frac{1}{|\mathcal{D}_i|} \sum_{(x_k, y_k) \subset \mathcal{D}_i} \ell\left(C_{\psi_i}\left(F_{o_i}(x_k)\right), y_k\right) + \lambda \frac{1}{|\mathcal{D}_i|} \sum_{(x_k, y_k) \subset \mathcal{D}_i} \left\| F_{o_i}(x_k) - \mathbf{H}_{y_k}^{r+1} \right\|_2^2 + \frac{\gamma}{n} \sum_{j=1}^{n} \left\| \frac{C_{\psi_i}\left(\mathbf{H}_j^{r+1}\right)}{T} - \frac{\mathbf{Q}_j^{r+1}}{T} \right\|_2^2 .$$

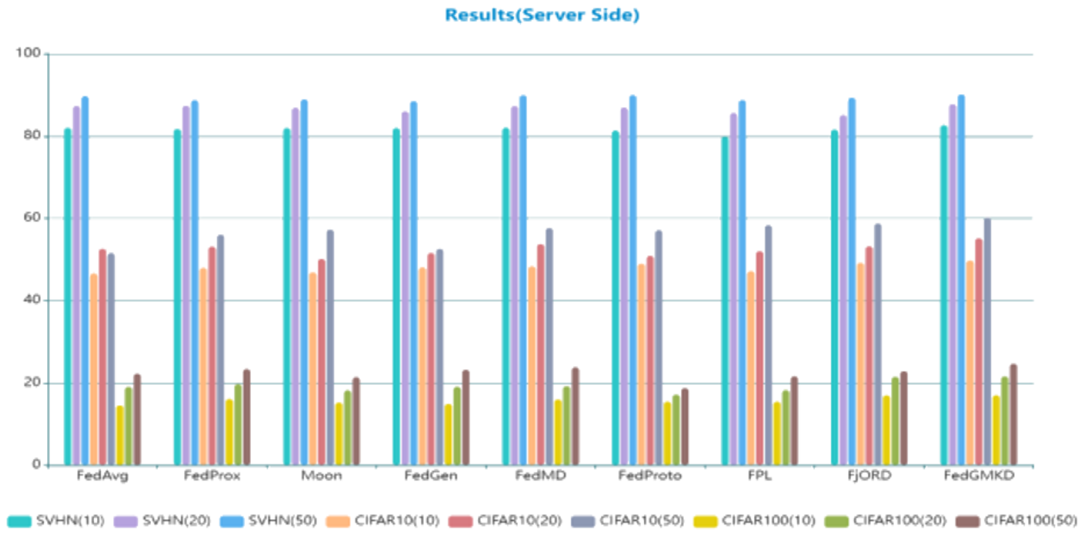
Experiments and Results

Experiment Settings

- Datasets: SVHN, CIFAR10, CIFAR100
- Model Architecture: ResNet-18 architecture
- Federated Setup:

a. Experiments conducted with varied numbers of clients (e.g., 10, 20, 50) to assess scalability
b. Each client trained for 3 local epochs per communication round
c. Total of 50 rounds to achieve convergence and assess global model performance

Dataset	Scheme	Local Acc			Global Acc			Avg Time (S)	Pub Data
		10	20	50	10	20	50		
SVHN	FedAvg	84.29	85.20	85.67	81.98	87.32	89.72	168.44	No
	FedProx	85.25	86.38	86.08	81.71	87.40	88.74	229.17	No
	Moon	84.11	85.43	85.43	81.95	86.90	88.97	358.14	No
	FedGen	85.18	85.10	84.96	81.96	86.02	88.52	205.37	No
	FedMD	85.45	85.90	86.31	82.04	87.30	89.91	611.33	Yes
	FedProto	85.58	86.44	86.85	81.34	86.97	89.79	346.13	No
	FPL	85.37	86.02	85.87	79.81	85.64	88.76	522.83	No
	FjORD	85.13	85.97	86.21	81.56	85.09	89.36	380.74	No
	FedGMKD	**86.26**	**87.43**	**87.16**	**82.64**	**87.78**	**90.17**	312.52	No
CIFAR10	FedAvg	55.75	58.76	61.51	46.62	52.61	51.53	98.94	No
	FedProx	57.46	58.91	62.94	47.97	53.13	56.04	126.56	No
	Moon	58.61	59.12	62.42	46.89	50.16	57.29	221.19	No
	FedGen	59.46	60.17	61.03	48.09	51.55	52.62	122.35	No
	FedMD	60.15	62.05	63.37	48.32	53.73	57.69	410.19	Yes
	FedProto	59.77	62.85	64.98	48.97	50.88	57.12	229.40	No
	FPL	60.95	62.74	64.49	47.19	52.04	58.35	295.97	No
	FjORD	59.62	63.36	63.61	49.18	53.22	58.74	252.34	No
	FedGMKD	**61.78**	**64.04**	**65.69**	**49.78**	**55.16**	**60.31**	251.55	No
CIFAR100	FedAvg	15.39	17.10	21.09	14.51	18.98	22.21	97.02	No
	FedProx	16.45	17.56	21.91	16.06	19.67	23.35	120.36	No
	Moon	15.46	18.03	21.25	15.19	18.16	21.37	201.91	No
	FedGen	14.08	17.05	19.54	14.88	19.05	23.16	148.58	No
	FedMD	13.25	19.03	21.93	15.96	19.20	23.75	482.76	Yes
	FedProto	15.70	18.63	22.50	15.38	17.13	18.72	206.12	No
	FPL	15.93	18.24	21.96	15.37	18.19	21.59	373.09	No
	FjORD	15.94	19.91	22.60	16.93	21.45	22.86	226.73	No
	FedGMKD	**17.16**	**20.96**	**23.57**	**16.97**	**21.56**	**24.63**	275.60	No



Impact of Model Complexity and Multi-Modality Performance

Scheme	ACC(Resnet18-local)	ACC(Resnet18-global)	ACC(Resnet50-local)	ACC(Resnet50-global)
FedAvg	61.78	49.78	41.69	49.58
FedProx	64.04	55.16	43.25	49.67
FedMD	62.05	53.73	43.34	49.85
FedGen	60.17	51.55	42.81	48.99
FedProto	62.85	50.88	43.35	49.98
Moon	62.74	52.04	42.05	48.52
FPL	62.74	52.04	43.71	49.78
FedGMKD	65.69	60.31	46.27	50.48

- FedGMKD demonstrates resilience with complex models like ResNet-50, leading all methods despite federated learning's communication and convergence constraints.

Scheme	ACC(local)	ACC(global)	Time(s)
FedAvg	83.71	50.52	411.95
FedProx	83.75	48.50	438.52
FedMD	83.87	48.29	700.73
FedGen	83.54	49.16	471.35
FedProto	84.13	49.75	586.77
FPL	83.96	50.12	665.29
FedGMKD	85.11	51.58	677.79

- FedGMKD is effective across data modalities, excelling in both vision and NLP tasks, highlighting its versatility in federated learning

FedGMKD Convergence Analysis

- **FedGMKD Convergence**

$$\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{n} w_i' \mathbb{E}\left[\|\nabla F_i(\mathbf{w}_i^r)\|^2\right] \leq \frac{F(\mathbf{W}^1) - F^*}{\eta R^2} + \sigma^2 + \frac{L\eta R G^2}{2}$$

- R : The total number of global communication rounds.
- n : The number of clients participating in federated learning.
- w_i' : The weight assigned to each client i, reflecting a combination of data quantity and data quality contributions.
- $\mathbb{E}\left[\|\nabla F_i(w_i^r)\|^2\right]$: The expected value of the squared gradient norm, representing the optimization state of the local model w_i.
- $F(W^1)$: The loss function value of the initial global model.
- F^* : The theoretical lower bound of the loss function.
- η : The learning rate.
- σ^2 : The variance of the gradient, which indicates uncertainty due to data heterogeneity during training.
- L: The Lipschitz constant, constraining the gradient's rate of change.
- G : The maximum gradient value, used to control the gradient's fluctuation range.
- **FedGMKD Convergence Rate**

$$F(\mathbf{W}^R) - F^* \leq \frac{C_1}{R} + C_2$$

- $F(W^R)$: The loss function value of the global model after R rounds.
- F^* : The theoretical lower bound of the loss function.
- C_1, C_2 : Constants dependent on gradient variance (σ^2), Lipschitz constant (L), learning rate (η), and the number of local steps.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you !