ClusterFL: A Similarity-Aware Federated Learning System for Human Activity Recognition

Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, Guoliang Xing
The Chinese University of Hong Kong

Presented by Rui Chen

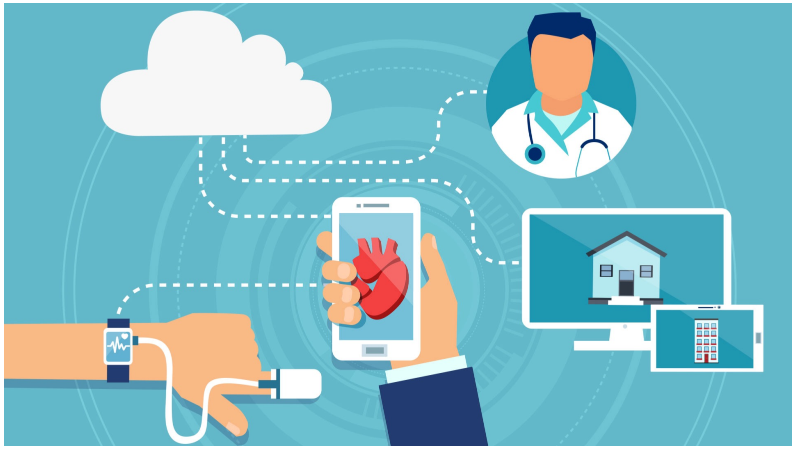
MobiSys 2021

Outline

- Motivation
- Similarity-aware federated learning framework
- Experimental Results
- Conclusion

Motivation – Human Activity Recognition

Fitness Tracking



Health monitoring

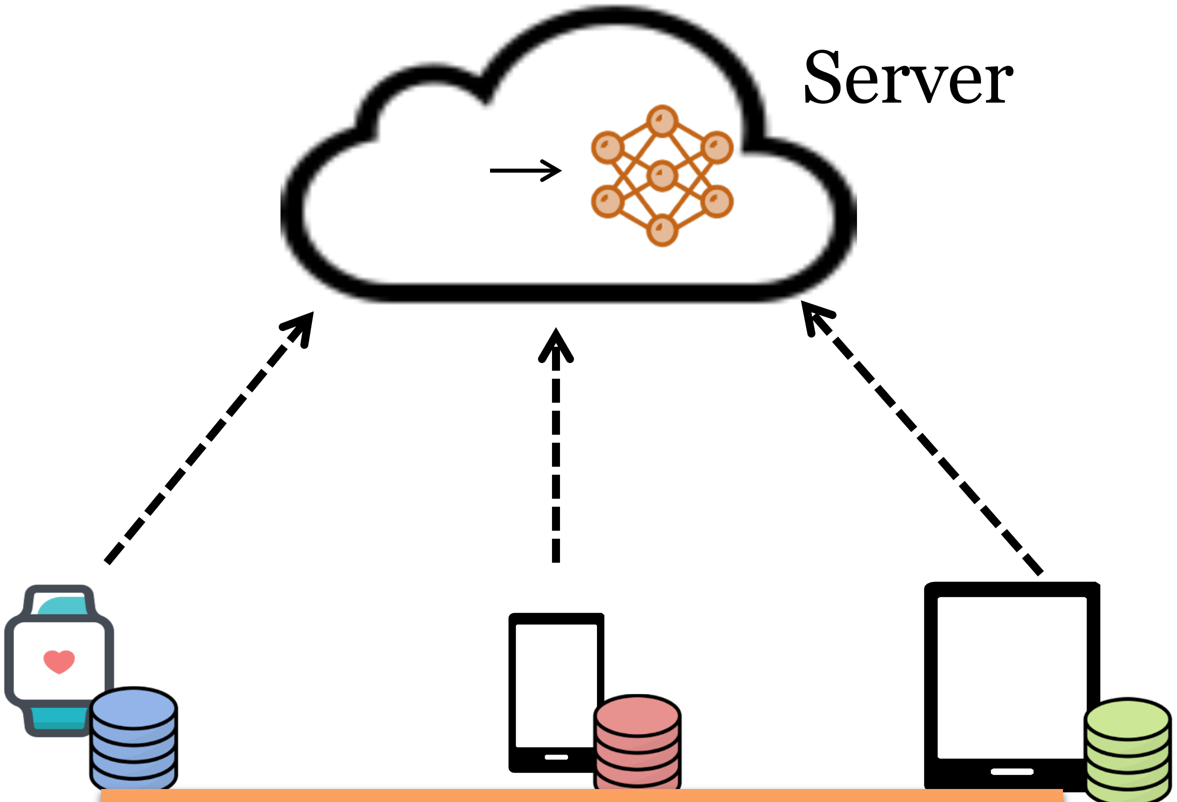
Sleep Monitoring

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation - Centralized Deep Learning

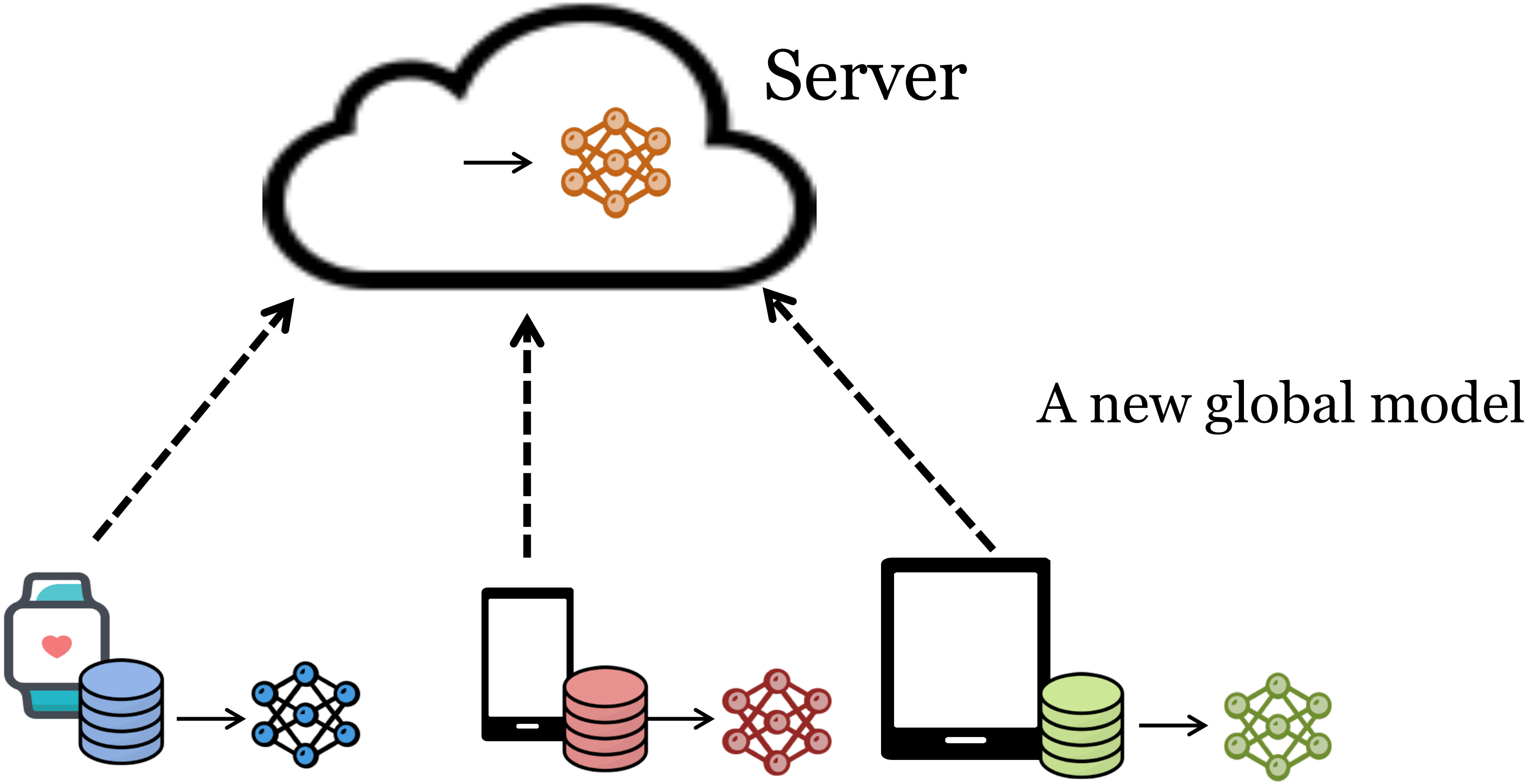
- Significant privacy concern (send raw data)



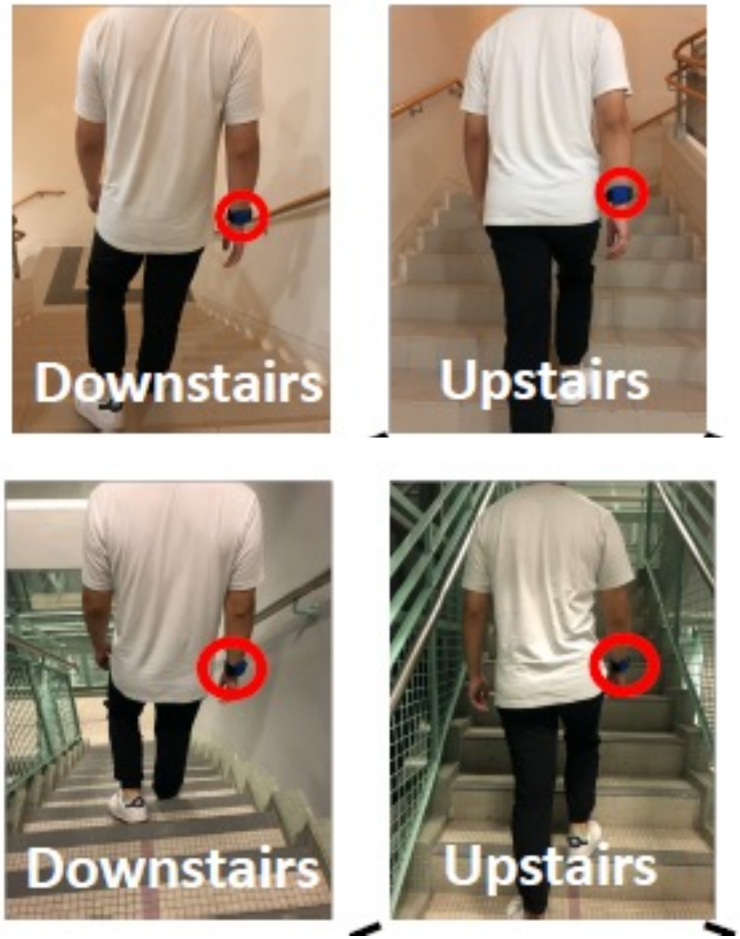
How about keep the data locally?

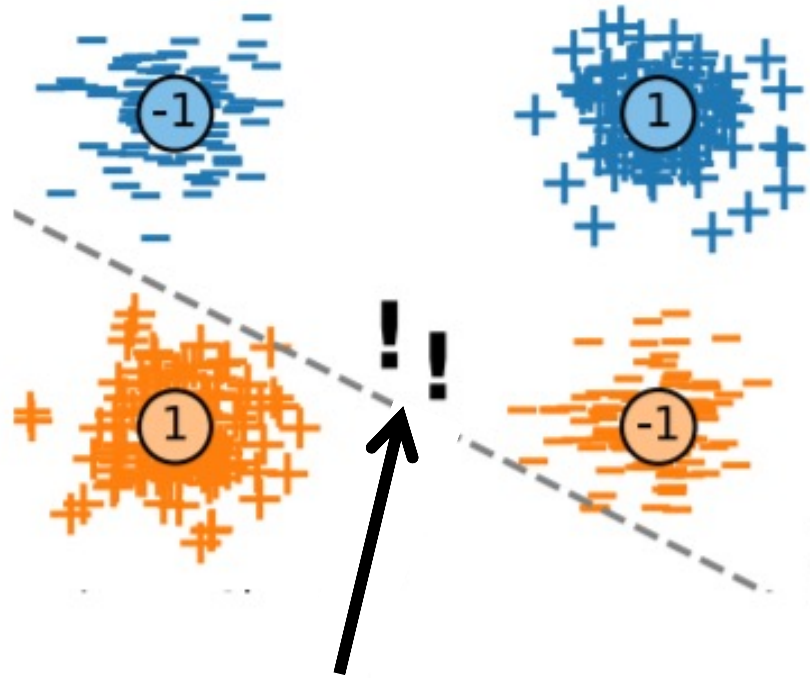
Motivation– Federated Learning

- Collaboratively training while keeping the data residing on devices.
- There is a high degree of data heterogeneity of human activity data.



Challenge – User Heterogeneity





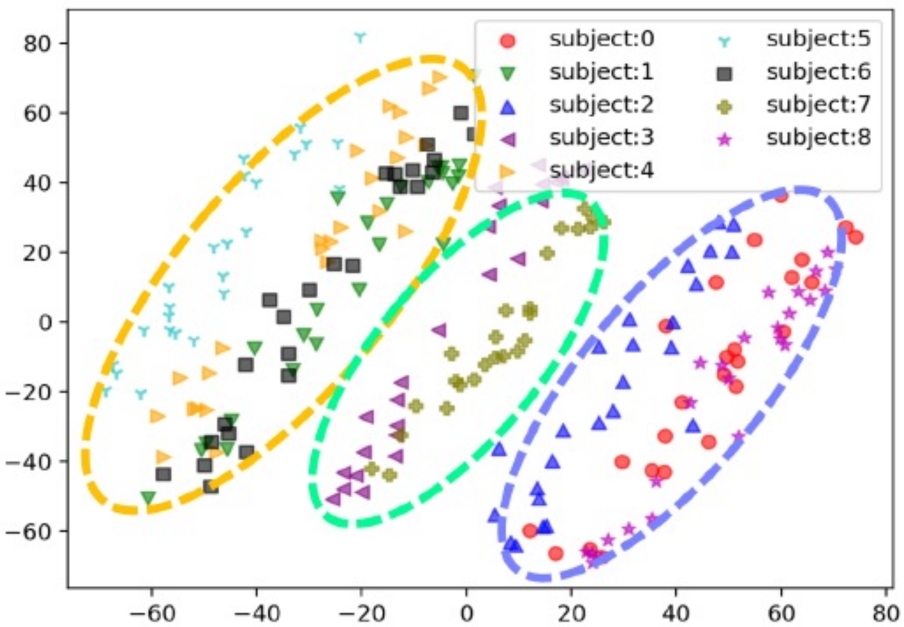
- In this situation, a single linear classifier can not simultaneously be correct for all clients.

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation

- There exhibits intrinsic similarities among users' activities.



due to

$\Longrightarrow$

Biological features

Physical environments

Sensor bias

A clear clustering relationship among subjects in the walking dataset, where the data from the same type of smartphone is grouped to the same cluster.

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation – Clusterability on Learning

Hopkins statistic: data X vs random data

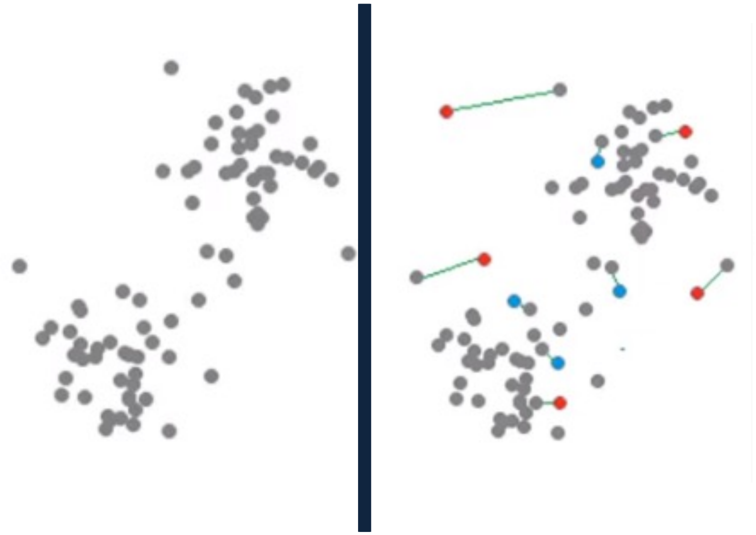
1. Blue data points ($p_1, \dots, p_m$) randomly selected from X;
2. Red data points ($q_1, \dots, q_m$) randomly generated from Uniform distribution.
3. Calculate distances:

   $\mathbf{d}_\mathrm{U}$: $q_1, \dots, q_m$ to the nearest points in X
   $\mathbf{d}_X$: $p_1, \dots, p_m$ to the nearest points in X

4. Hopkins statistic

$$\boldsymbol{H} = \frac{\sum_{i=1}^{m} \mathbf{d}_{\mathrm{U},i}}{\sum_{i=1}^{m} \mathbf{d}_{\mathrm{X},i} + \sum_{i=1}^{m} \mathbf{d}_{\mathrm{U},i}}$$

$H \approx 1$: data highly clustered
$H = 0.5$: data random
$H \approx 0$: data uniform



data X

Motivation – Clusterability on Learning

Empirical studies

- Quantifies the clustering tendency of HAR data

Dataset	SHAR	HHAR	Depth	IMU	UWB	HARBox
# of subjects/activities	30/6	9/6	9/3	7/2	8/3	121/5
Hopkins statistic	0.8813	0.7951	0.8699	0.6966	0.5742	0.8946

- The Centralized-cluster method has the bes

Grouping helps!

Method	Local	FedAvg	FTL	Centralized-single	Centralized-cluster
Mean Accuracy (%)	53.67	33.61	54.61	72.22	73.17
STD (%)	9.74	12.812	10.19	6.06	3.40

Nodes share a global model

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Contribution

- propose ClusterFL, a similarity-aware federated learning system to enable collaborative learning among similar nodes.

- introduce two communication efficient strategies: cluster-wise straggler dropout and correlation based node selection, based on the learned cluster structure.

- collect four new human activity recognition (HAR) datasets in different environments with significant dynamics

Outline

- Motivation
- Similarity-aware federated learning framework
- Experimental Results
- Conclusion

Similarity-aware FL Framework

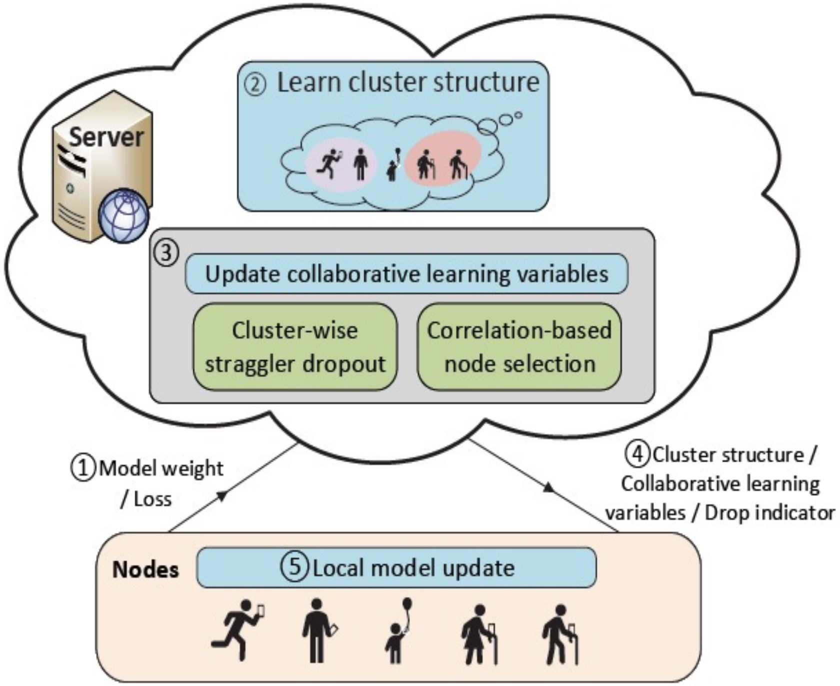


Figure 2: System Architecture of ClusterFL

Step 1: Nodes upload local model weights/loss

Step 2: The server use a **cluster indicator matrix** to quantify the clusters

Step 3: Based on the learned cluster indicator matrix, the server updates the models and **identify cluster-wise slow/less important nodes**

Step 4: The updated models and a drop indicator are send back to nodes

Step5: Nodes perform the local training

Similarity-aware FL Framework

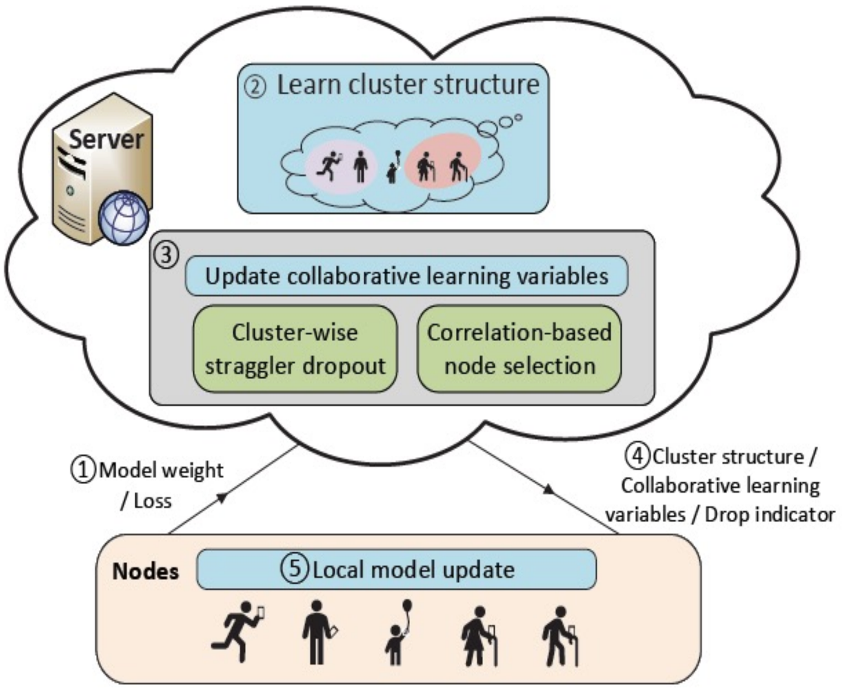


Figure 2: System Architecture of ClusterFL

- **Similarity-aware FL framework**
  - How to identify clusters
  - How to update model

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Similarity-aware FL Framework

- The *global objective* of ClusterFL

cluster indicator matrix

Local model

$\mathbf{F} \in \mathbb{R}^{M\times K}$

$$\min_{\mathbf{W},\mathbf{F}} \sum_{i=1}^{M} \frac{1}{N_i} \sum_{r=1}^{N_i} l(\mathbf{w}_i^T \mathbf{x}_i^r, y_i^r) + \underbrace{\alpha tr(\mathbf{W}\mathbf{W}^T) - \beta tr(\mathbf{F}^T \mathbf{W}\mathbf{W}^T \mathbf{F})}$$

$[\mathbf{w}_1, ..., \mathbf{w}_M]^T \in \mathbb{R}^{M\times D}$

$$\boxed{(\alpha - \beta) \sum_{i=1}^{M} ||\mathbf{w}_i||_2^2 + \beta \sum_{j=1}^{K} \sum_{v \in \mathcal{S}_j} ||\mathbf{w}_v - \overline{\mathbf{w}_j}||_2^2,}$$

Customize a model for each node

L2-Regularization

K-means clustering

Learn from the same cluster

M: the number of nodes
N_i: the number of training samples at node i's dataset
α, β: hyperparameters

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Similarity-aware FL – Alternating optimization

x_0^r : the public available data at server sides.

τ : parameter to adjust the sensitivity of the model outputs

- Fixed **W**, optimize **F**

  $\longrightarrow$ via the similarity of the node models

$$D_{KL}(\mathbf{w}_i, \mathbf{w}_j) = \frac{1}{N_O} \sum_{r=1}^{N_O} \delta(\mathbf{w}_i, \mathbf{x}_o^r) log \frac{\delta(\mathbf{w}_i, \mathbf{x}_o^r)}{\delta(\mathbf{w}_j, \mathbf{x}_o^r)}$$

$$\delta(\mathbf{w}_i, \mathbf{x}_o^r) = softmax(\frac{\Phi(\mathbf{w}_i, \mathbf{x}_o^r)}{\tau})$$

pre-softmax output of model

- K-means clustering $\mathbf{F} \in \mathbb{R}^{M \times K}$

Unknown

$F_{i,j} = \frac{1}{\sqrt{N_j}}$ if node i belongs to j-th cluster and $Fi, j = 0$ otherwise

ANTS LAB

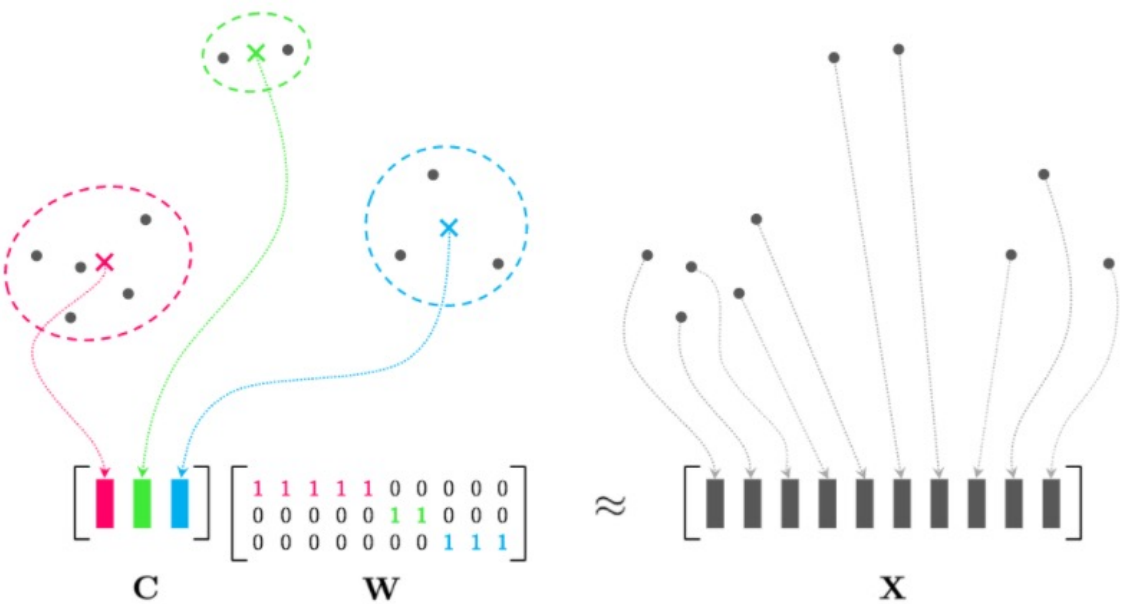
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Similarity-aware FL -- Cluster indicator

- *K*-means clustering

Unknown number of clusters K

Solution[1]: principal components analysis (PCA)



[1] Ding et al. K-means clustering via principal component analysis. 2004

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Similarity-aware FL -- Cluster indicator

- Fixed **W**, optimize **F**

Solution[1]: Based on the KL divergence matrix $\mathbf{D} \in \mathbb{R}^{M \times M}$

M-1 principal components

Normalized indicators F

$$\mathbf{P} = Q_{M-1} Q_{M-1}^{T} \in \mathbb{R}^{M \times M} \implies F_{ij} = \frac{P_{ij}}{\sqrt{\sum_{i=1}^{M} P_{ij}}}$$

$(\text{set } P_{ij} = 0 \text{ if } P_{ij} < 0)$

P is the continuous solution of K-means clustering to approximate **F.**[1]

[1] Ding et al. K-means clustering via principal component analysis. 2004

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Similarity-aware FL – Alternating optimization

- Fixed **F**, optimize **W**, **Ω**

$$\min_{\mathbf{W},\Omega} f(\mathbf{W}) + g(\Omega)$$
$$s.t.\ \mathbf{F}^T\mathbf{W} - \Omega = 0$$
$$\text{where: } f(\mathbf{W}) = \sum_{i=1}^{M} \frac{1}{N_i} \sum_{r=1}^{N_i} l_t(\mathbf{w}_i^T \mathbf{x}_i^r, y_i^r) + \alpha tr(\mathbf{W}\mathbf{W}^T)$$
$$g(\Omega) = -\beta tr(\Omega\Omega^T)$$

- node update **W**:

$$\mathbf{w}_i^{t+1} = \arg\min_{\mathbf{w}_i} \frac{1}{N_i} \sum_{r=1}^{N_i} l_t(\mathbf{w}_i^T \mathbf{x}_i^r, y_i^r) + (\alpha + \frac{\rho}{2} \sum_{j=1}^{K} F_{ij}^2)||\mathbf{w}_i||_2^2 + \sum_{j=1}^{K} F_{ij}(\mathbf{U}_j^t - \rho\Omega_j^t)^T \cdot \mathbf{w}_i$$

- server update **U**, **Ω**:

$$\Omega_j^{t+1} = \arg\min_{\Omega_j} -\beta||\Omega_j||_2^2 - \Omega_j \cdot (\mathbf{U}_j^t)^T + \frac{\rho}{2}||\Omega_j - \mathbf{F}_j^T\mathbf{W}^{t+1}||_2^2$$
$$\mathbf{U}_j^{t+1} = \mathbf{U}_j^t + \rho(\mathbf{F}_j^T\mathbf{W}^{t+1} - \Omega_j^{t+1})$$

Similarity-aware FL

Algorithm 1: Alternating Optimization of ClusterFL

Initialization : set $\mathbf{W} = 0^{M \times D}$, $\mathbf{F} = \mathbf{0}^{M \times M}$

for $h = 0$ *to* H **do**

  1. Optimization of **W** with **F** fixed.

  for $t = 0$ *to* T **do**

    node update: parallelly update $\mathbf{w}_i$(i = 1,...,M)

      $\mathbf{w}_i^{t+1} \leftarrow \textbf{Local SGD}(\mathbf{w}_i^t, (\mathbf{x}_i^r, y_i^r), \mathrm{F}, \Omega, \mathrm{U})$

    server update: update Ω and U

    for $j = 1$ *to* K **do**

      $\Omega_j^{t+1} \leftarrow \arg\min_{\Omega_j} \mathrm{L}_\rho(\Omega_\mathrm{j}, \mathrm{U}_j^t, \mathrm{F}_j^t, \mathrm{W}^{t+1})$

      $\mathrm{U}_j^{t+1} \leftarrow \mathrm{U}_j^t + \rho(\mathrm{F}_j \mathrm{W}^{t+1} - \Omega_j^{t+1})$

    end

  end

  2. Optimization of **F** with **W** fixed.

  server update: Update **F** based on Section 5.3

end

Similarity-aware FL Framework

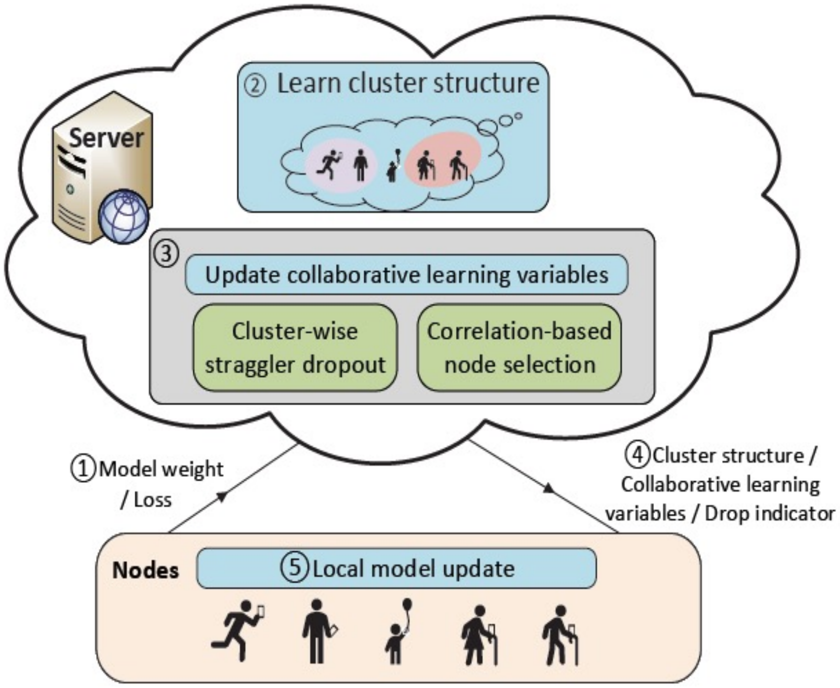


Figure 2: System Architecture of ClusterFL

- **Similarity-aware FL framework**
  - How to identify clusters
  - How to update model
- **Communication Efficiency**
  - How to handle straggler
  - How to identify "important" nodes

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Similarity-aware FL Framework

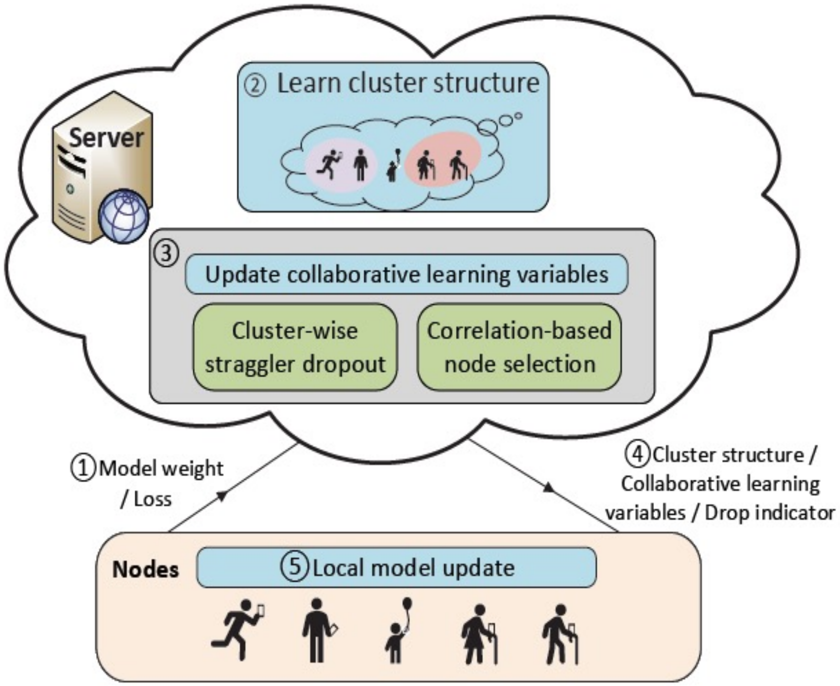


Figure 2: System Architecture of ClusterFL

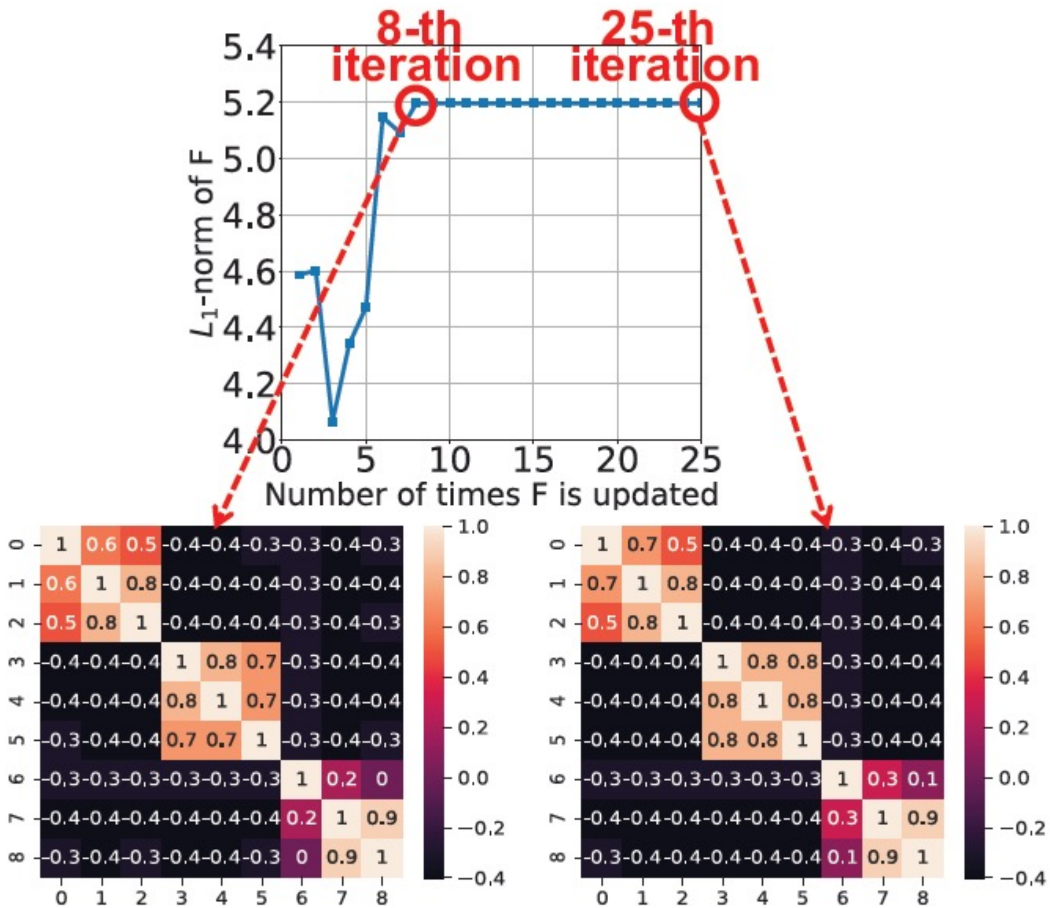
- **Communication Efficiency**
  - How to handle straggler
  - How to identify "important" nodes

Idea: leverage cluster structure to drop the nodes while keep the similar accuracy

Optimizing Communication Performance

The cluster relationship can be learned in the early stage.

F



Optimizing Communication Performance

✓ Drop the cluster-wise stragglers (the nodes who converge slower)

According to cluster structure **F**, we can measure the cluster-wise convergence performance:

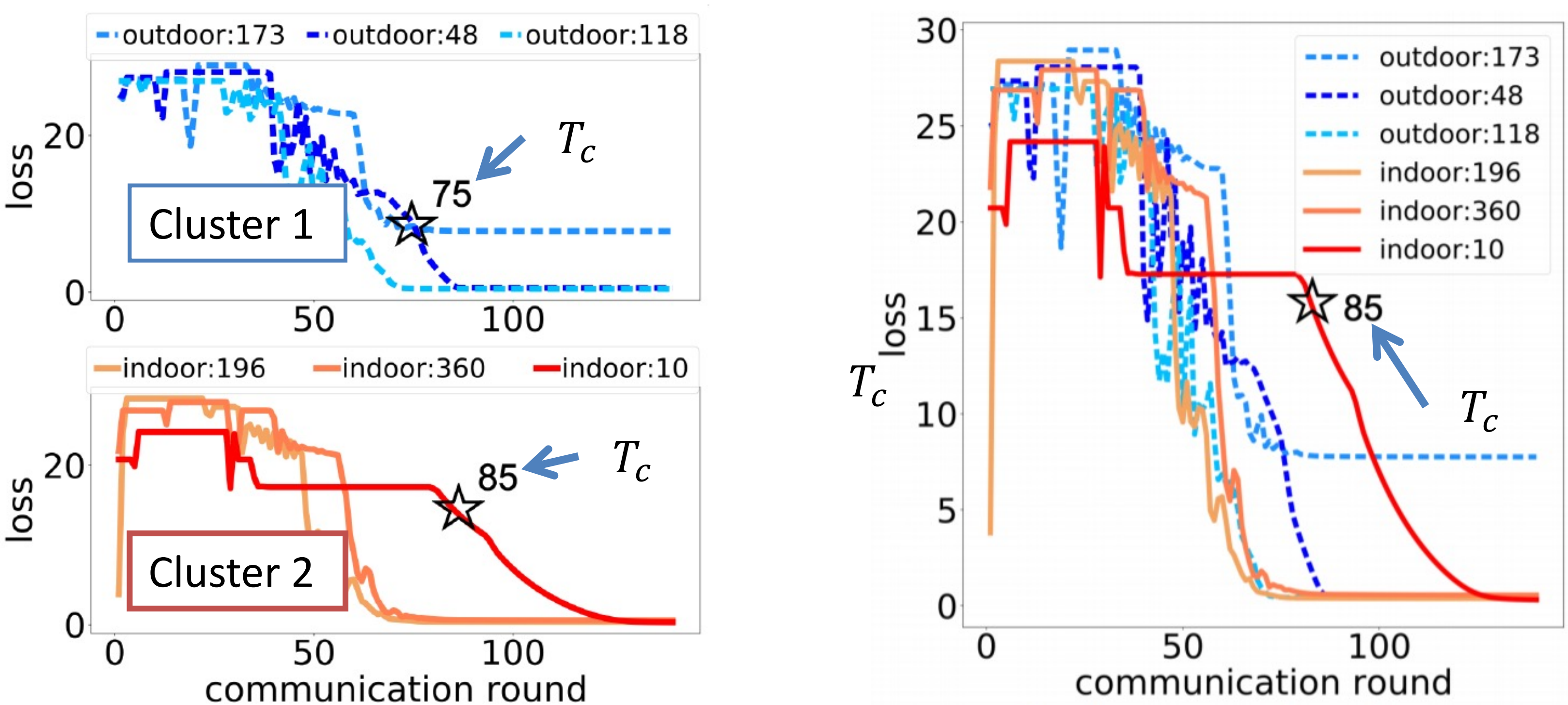
$$\varepsilon_q^t = \frac{|loss_q(t) - loss_q(t-1)|}{\sum_{q=1}^{N_j} |loss_q(t) - loss_q(t-1)|}$$

$$\gamma_q = \frac{1}{T_c} \sum_{t=1}^{T_c} \varepsilon_q^t \quad for\ q = 1, 2, \dots, N_j$$

T_c: the threshold of iteration for determining stragglers at each cluster

Optimizing Communication Performance

✓ Drop the cluster-wise stragglers (the nodes who converge slower)



(a) Within clusters

(b) Among all nodes

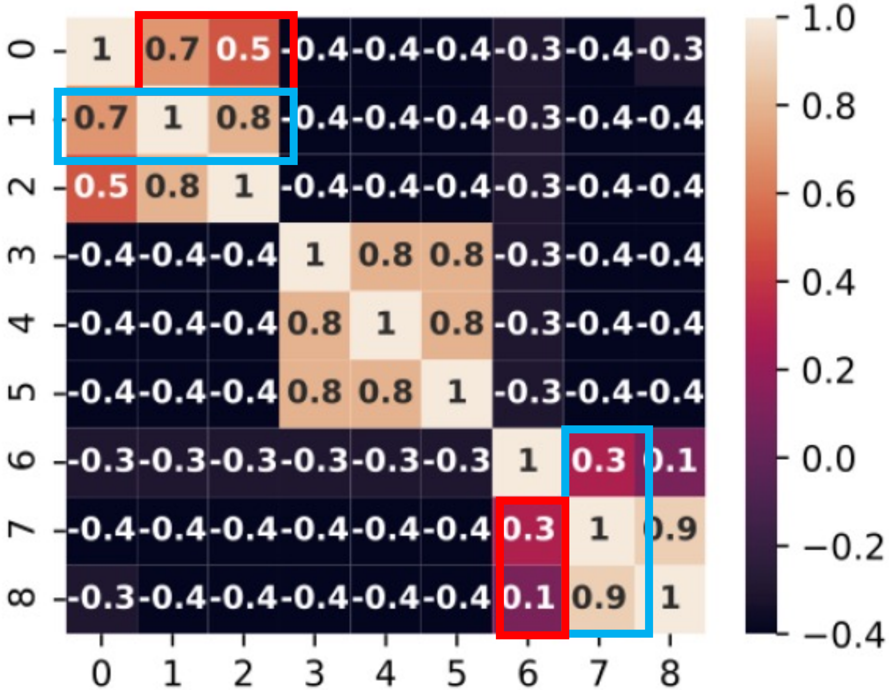
More dynamic and flexible

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

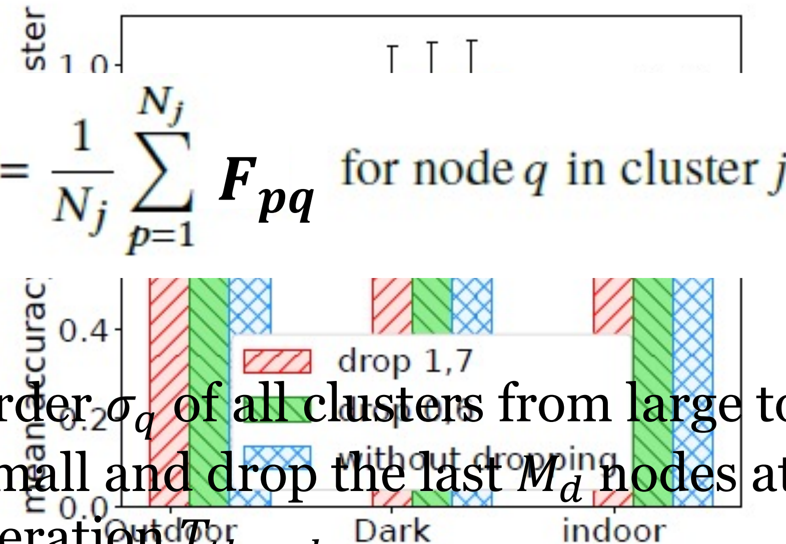
Optimizing Communication Performance

✓ Drop nodes whose data is less related to other nodes in the same cluster.

- Node 0 has weaker correlations in the first cluster. (Outdoor)
- Node 1 has stronger correlations in the first cluster.



$$\sigma_q = \frac{1}{N_j} \sum_{p=1}^{N_j} \boldsymbol{F}_{pq} \quad \text{for node } q \text{ in cluster } j$$



order σ_q of all clusters from large to small and drop the last M_d nodes at iteration T_{thresh}.

(b) Mean accuracy of each cluster

Outline

- Motivation
- Similarity-aware federated learning framework
- Experimental Results
- Conclusion

Experiment Setup

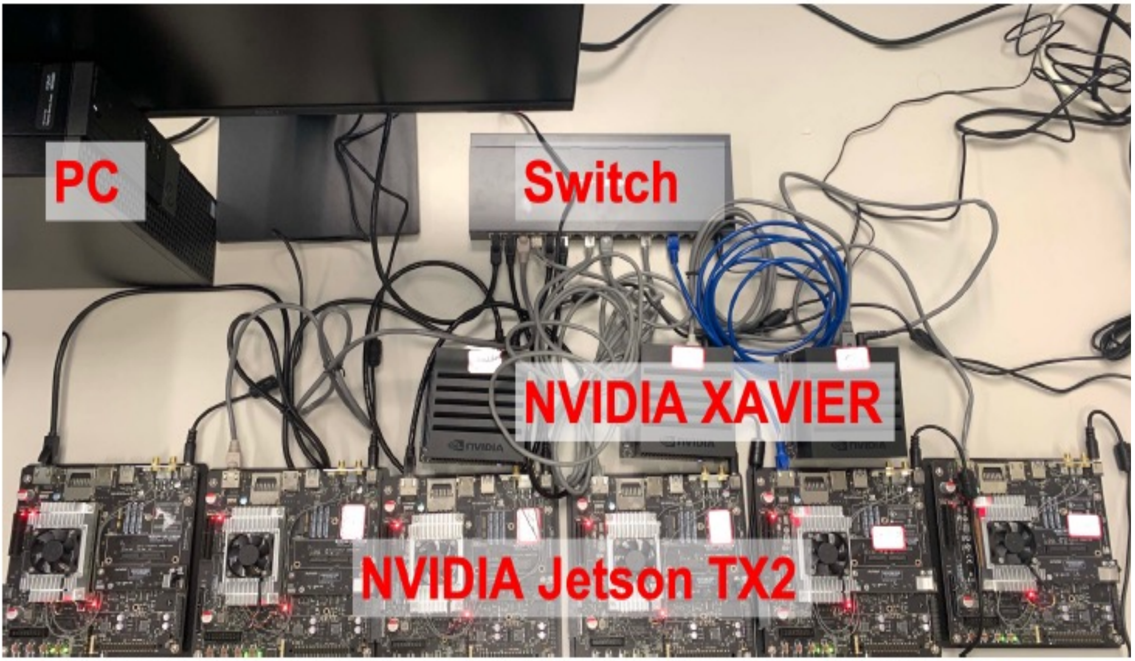


Figure 10: Hardware Setup

- The server and nodes are connected via a TP-link TL-SG2016K switch.
- The power are measured via the on-board power monitoring sensor TIINA3221x.

Experiment Setup

- **Dataset**: four new human activity recognition (HAR) datasets

Application	Task	Data Dimension	Number of Subjects	Number of Data Records	Sensor	Environment
Human Movement Detection using UWB	with/without Human Movement	55	8	663	Decawave DWM1000 UWB	node 0,1 from parking lot node 2,3,4 from corridor node 5,6,7 from room
Walking Activity Recognition using IMU	walking on corridors/ upstairs/downstairs	900	7	1369	LPMS-B2 IMU	node 0,1,2,3 from building 1 node 4,5,6 from building 2
Gesture Recognition using Depth Camera	good/ok/ victory/stop/fist	1296	9	7422	PicoZense DCAM710	node 0,1,2 from "outdoor" node 3,4,5 from "dark" node 6,7,8 from "indoor"
HARBox: ADL Recognition using Smartphones	walking/hopping/ phone calls/waving/typing	900	121	32935	77 different smartphone models	121 subjects (17-55 years old) Sampling rate: 43.5-57.5Hz

- **Models**: SVM for human movement detection;
  Five-layer NN for walking activity recognition;
  Five-layer CNN for depth camera gesture recognition.

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

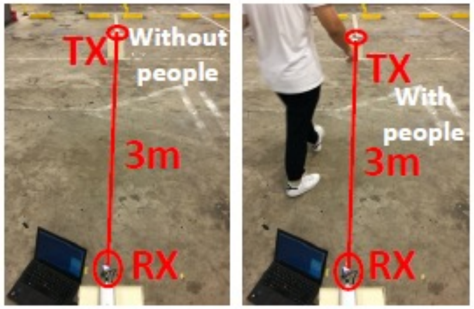
Experiment Setup

- **Baseline**:
  1) Local training
  2) Federated average[1] (FedAvg)
  3) Federated transfer learning[2] (FTL)
  4) Centralize (single) model training

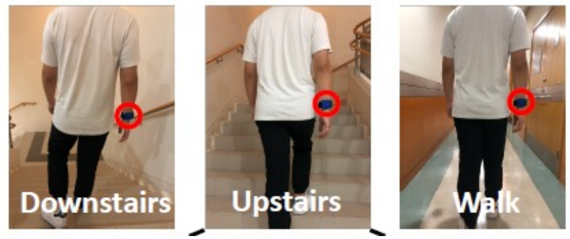
[1] McMahan Communication-Efficient Learning of Deep Networks from Decentralized Data, 2017

[2] Feng et alPmf: A privacy-preserving human mobility prediction framework via federated learning. 2019

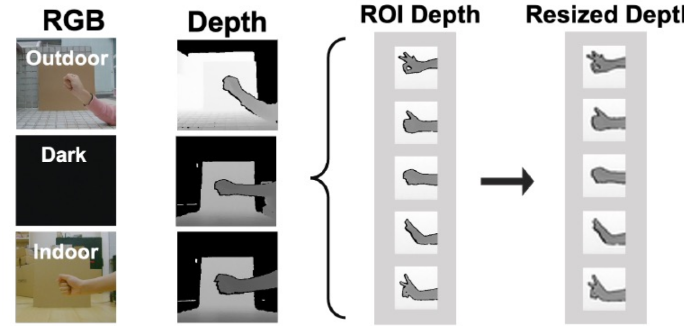
Visualization of Different Datasets



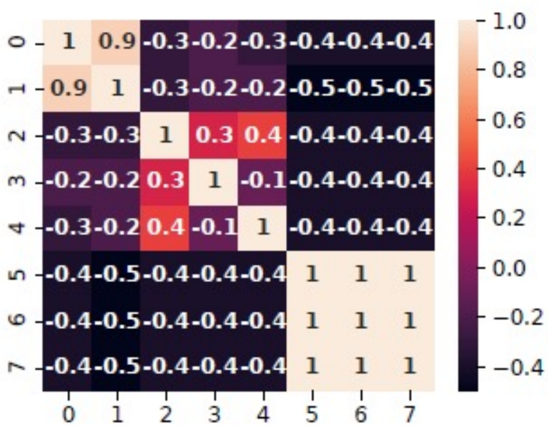
multi-path effect of
Ultra Wide Band signals



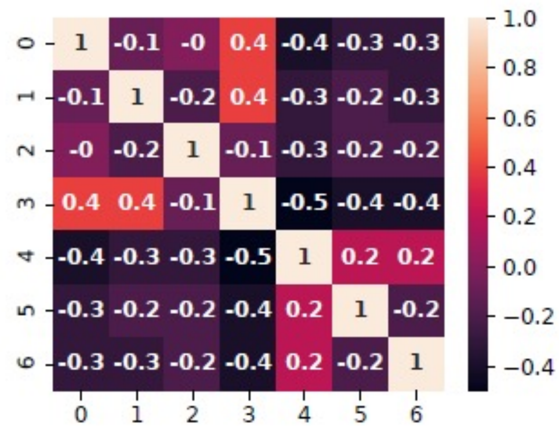
Sensor data from
smart watch



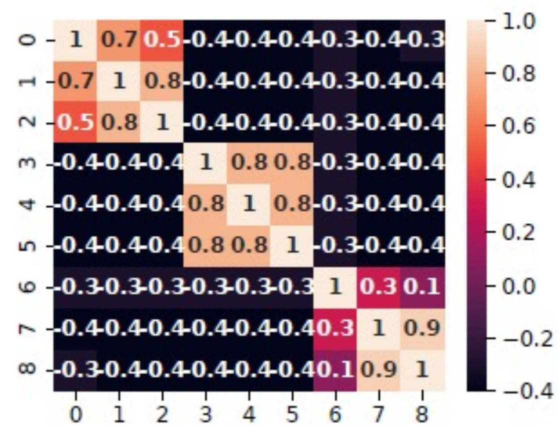
Camera image data



(a) UWB structure



(b) IMU structure

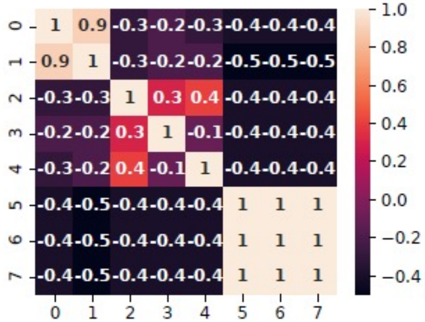


(c) Depth structure

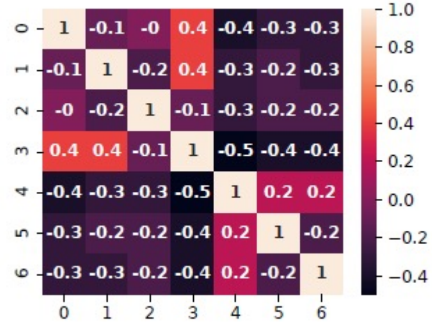
ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

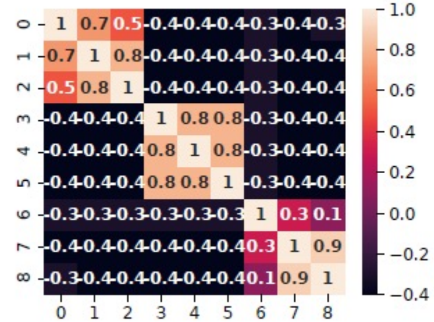
Performance on Different Datasets



(a) UWB structure



(b) IMU structure



(c) Depth structure

	Local	FedAvg	FTL	Centralized	ClusterFL
UWB	71.67%	86.25%	87.30%	**92.71%**	**92.71%**
IMU	88.29%	82.00%	86.20%	84.19%	**89.05%**
Depth	67.91 %	63.75%	63.86%	**96.82%**	70.68%

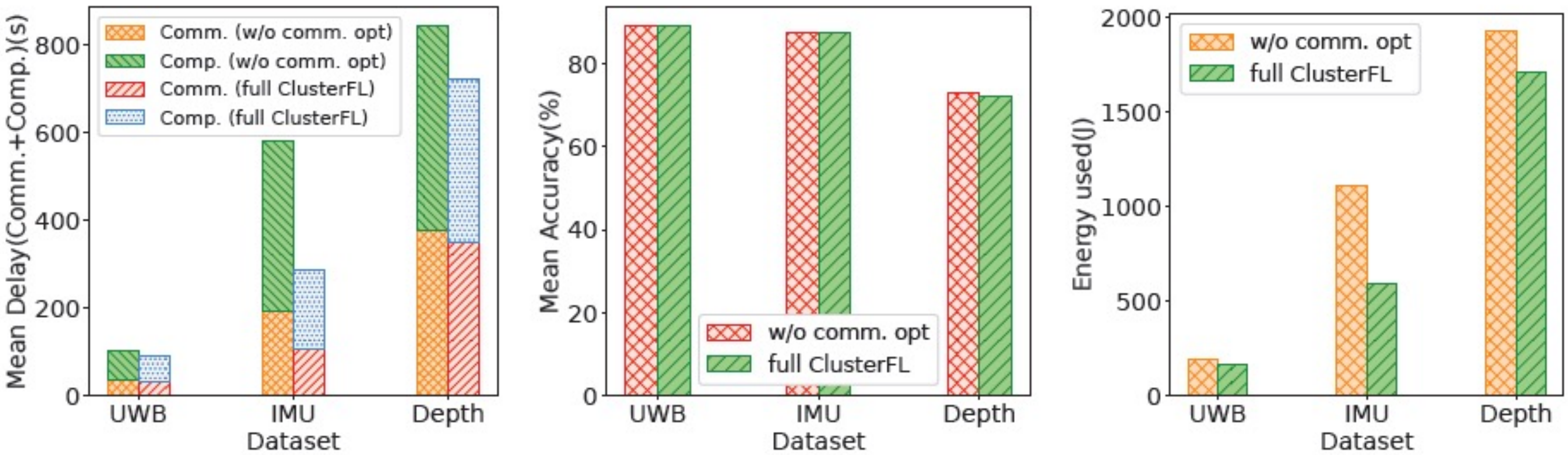
Table 5: Mean accuracy in unbalanced data settings

- ClusterFL outperforms Local/FedAvg/FTL in model accuracy.
- For the IMU dataset, ClusterFL has the best performance in model accuracy.

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Performance on Different Datasets

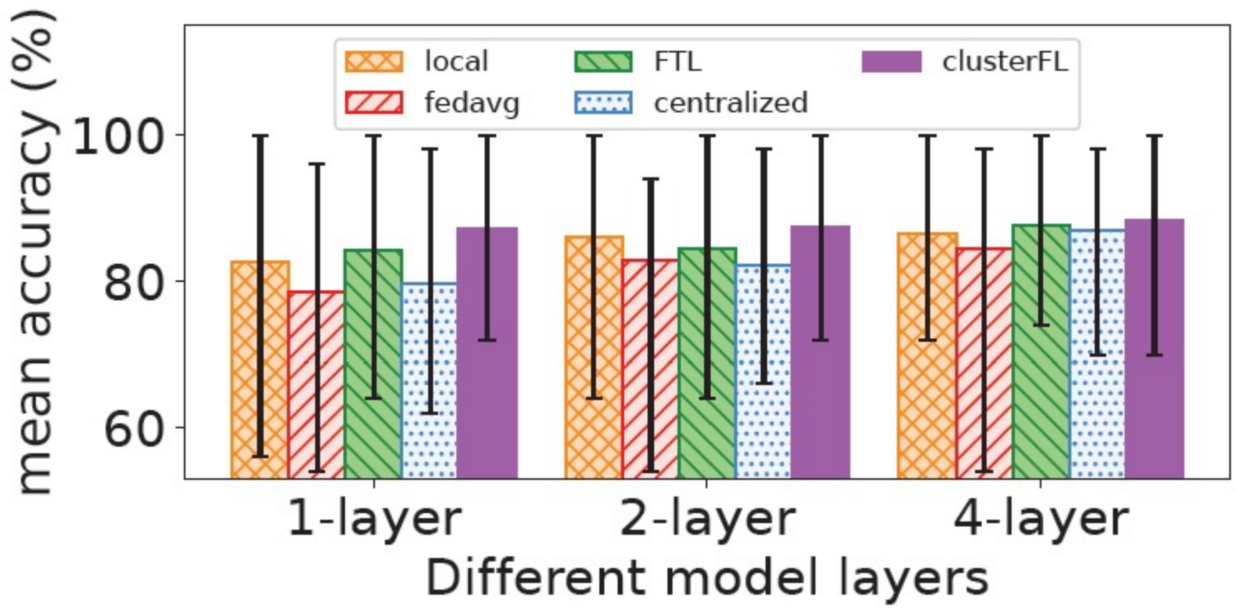


(a) Time delay **(b) Accuracy** **(c) Energy**

- ClusterFL can reduce more than 20% communication latency while maintaining almost the same accuracy performance.

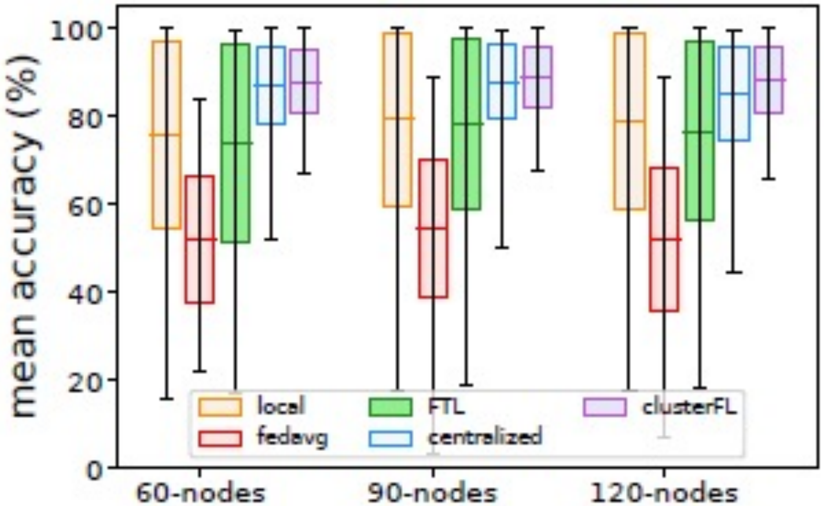
Performance on Model Depths

IMU dataset



- When the model becomes deeper, the accuracy of all methods tends to increase.
- ClusterFL has a more “stable" accuracy performance under different learning models

Performance on the scalability of ClusterFL



(a) Mean accuracy

(b) Mean commu. cost per node

- When the number of nodes increases, the mean accuracy of centralized learning slightly decreases, suggesting the heterogeneity of user' data.
- ClusterFL has a significantly smaller variation of accuracy among nodes and save more than 50% communication time.

Conclusion

- ClusterFL features a novel federated learning framework enabling collaborative learning among similar nodes

- ClusterFL integrates two effective communication optimization mechanisms based on the learned cluster structure.

- ClusterFL outperforms several learning paradigms and reduce more than 50% communication latency


THANK YOU
UNIVERSITY of HOUSTON | ENGINEERING