Towards Collaborative Multi-modal Federated Learning Human Activity Recognition in Smart Workplace Enviornments

UNIVERSITY OF TORONTO

SM. Sheikholeslami, P. C. Ng, H. Liu, Y. Yu, K. N . Plataniotis

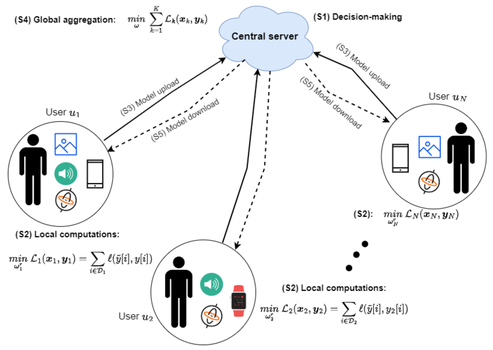
Department of Electrical and Computer Engineering

Introduction

- **Data over networks comes**
  - **in form of different modalities**
  - **Is distributed over multiple devices**

- improve **human activity recognition (HAR)** with multimodal data across multiple consumer devices in a smart workplace environment.

- Preserve privacy of users by **Multi-modal Federated Learning** to collaboratively train a global model



Research question 1) How can data distributed over multiple devices of a single user be combined to create a super model?

Research question 2) How can private data from different user can be combined to train a richer super model?

Contributions:

1) Local (Intra-zone) training: Collaboration of multiple devices with different sensing inputs belonging to the same user to train a local multimodal HAR model

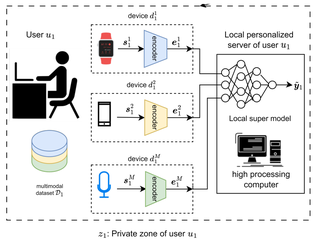
2) Global (Inter-zone) training: the cross-user cooperation of different users enriches the global model without sharing their private multimodal data

3) A robust feature reconstruction network to adaptively reconstruct the missing data.

4) A comparative analysis between unimodal and multimodal approaches within both centralized and distributed settings with missing features

Methods

Private zone z_k: The set of devices belonging to each user u_k



Proposed fusion model

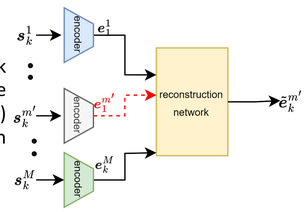
a) Feature extraction: At user u_k, a distinct encoders is assigned for each modality m to map the vectorized sensory inputs $s_m^k \in \mathbb{R}^{S_m^k}$ to a feature representation $e_m^k \in \mathbb{R}^{e_m}$

b) Multi-modal feature fusion: Subsequently, concatenated feature vector $e_k = [e_1^k, \dots, e_M^k]$ serves as the input for the activity classifier c_k.

c) Feature reconstruction: Let m' denote the missing modality in the inference phase. This missing modality is reconstructed using available data modalities by minimizing

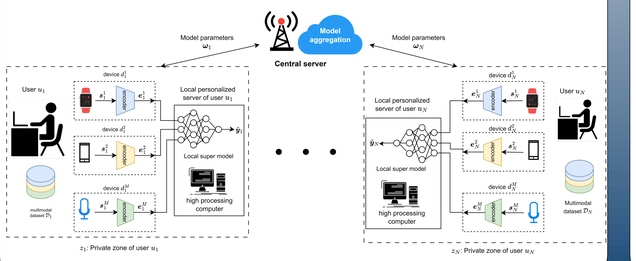
$$\mathcal{L}_k^{MR} = \sum_{i=1}^{|\mathcal{D}_{u_k}|} L(\tilde{e}_k^{m'}[i], e_k^{m'}[i])$$

Modality reconstruction network to reconstruct the feature (indicated by the red dashed line) lost due to the communication failure.



Proposed Collaborative Multi-modal Federated Learning (CoMFL)

Our proposed CoMFL consists of Intra-zone for local collaborative training between local devices enabling efficient processing of multiple modalities from multiple devices, and Inter-zone for facilitating global federated learning,



Results

A) Simulation Setup:

1) Dataset:

- **Total 28 users:**
  - **20 active users: participate in training**
  - **8 passive users: do not participate in training**
- **Each user possesses**
  - Smartphone
  - Smartwatch
  - Smartspeaker
  - Computer

Table 1: List of activities in the multimodal HAR dataset.

Simple activities	Complex activities
(a1) walking around	**(a8)** walking and talking on the phone
(a2) talking on the phone	
(a3) reading book	
(a4) stretching	
(a5) writing	**(a9)** writing and talking on the phone
(a7) browsing the monitor	
(a6) typing on the keyboard	

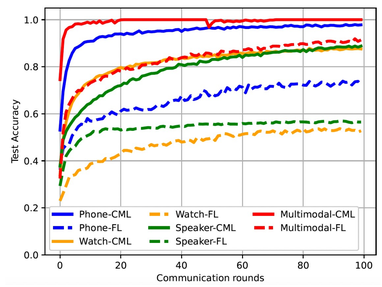
2) Multimodal Deep Learning Model:

- For each device, we use 2 CNN layers with 64 and 32 and filters of size 3x3 and max-pooling 2x2. The super model is composed of two fully-connected layers with 64 and 32 neurons, respectively.

B) Experimental results:

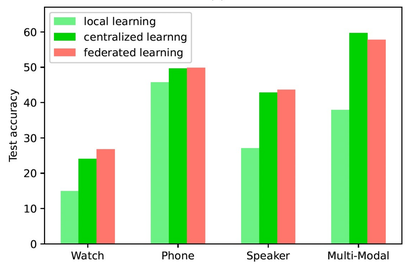
Experiment 1)
Test accuracy vs Communication rounds (epochs) for active users

- Phone modality provides the best accuracy among unimodal HAR models
- In general, central training outperforms federated learning
- Multi-modal fusion enhances performance of HAR models



Experiment 2)
Performance on cross-domain data (passive users)

- The propose CoMFL enhances generalization
- CoMFL performs almost same as centralized case while preserving the privacy of users



Experiment 3)
Extracted features could be missed due to communication failure

- Phone modality has the greatest effect on the performance of the super model when missing
- The proposed reconstruction model recovers the missing modality and enhances performance of the model

Table 2: Performance of the multimodal HAR models with missing modality.

training strategy \ modality	missing modality						full modality
	watch		phone		speaker		
	baseline	proposed	baseline	proposed	baseline	proposed	
MM centralized learning model	83.21%	90.83%	38.11%	46.17%	94.54%	98.14%	99.29%
MM federated learning model	74.03%	78.44%	28.33%	54.26%	79.21%	85.79%	91.37%

Conclusions

- The proposed CoMFL framework enables privacy-preserving feature encoding and fusion by efficiently leveraging the sensor capabilities of smartphones, smartwatches, and smart speakers,
- The incorporation of federated learning and feature reconstruction further enhances performance, even in the presence of missing data.
- CoMFL demonstrates significant improvements in multimodal HAR systems, showcasing its potential for future applications in privacy-preserving, smart workplace environments.

Bibliography

1. Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu, "Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 36, no. 3, pp. 3063–3071, Jun. 2022.
2. Dae Yon Hwang, Pai Chet Ng, Yuanhao Yu, Yang Wang, Petros Spachos, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis, "Hierarchical deep learning model with inertial and physiological sensors fusion for wearable-based human activity recognition," *in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 21–25, 2022