Federated Multi-Task Attention for Cross-Individual Human Activity Recognition

Qiang Shen[1] , **Haotian Feng**[1] , **Rui Song**[2] , **Stefano Teso**[3] , **Fausto Giunchiglia**[1,3] and **Hao Xu**[1,2*]
[1]College of Computer Science and Technology, Jilin University
[2]School of Artificial Intelligence, Jilin University
[3]University of Trento
{shenqiang19, fenght21, songrui20}@mails.jlu.edu.cn, {fausto.giunchiglia, stefano.teso}@unitn.it, xuhao@jlu.edu.cn

Abstract

Federated Learning (FL) is an emerging privacy-aware machine learning technique that applies successfully to the collaborative learning of global models for Human Activity Recognition (HAR). As of now, the applications of FL for HAR assume that the data associated with diverse individuals follow the same distribution. However, this assumption is impractical in real-world scenarios where the same activity is frequently performed differently by different individuals. To tackle this issue, we propose **FedMAT**, a **Fed**erated **M**ulti-task **AT**tention framework for HAR, which extracts and fuses shared as well as individual-specific multi-modal sensor data features. Specifically, we treat the HAR problem associated with each individual as a different task and train a federated multi-task model, composed of a shared feature representation network in a central server plus multiple individual-specific networks with attention modules stored in decentralized nodes. In this architecture, the attention module operates as a mask that allows to learn individual-specific features from the global model, whilst simultaneously allowing for features to be shared among different individuals. We conduct extensive experiments based on publicly available HAR datasets, which are collected in both controlled environments and real-world scenarios. Numeric results verify that our proposed **FedMAT** significantly outperforms baselines not only in generalizing to existing individuals but also in adapting to new individuals.

1 Introduction

Human activity recognition (HAR) plays an important role in context-aware services such as health monitoring and aging care [Straczkiewicz *et al.*, 2021]. HAR involves collecting and processing personal behavior data for training purposes, which has important consequences in terms of data privacy. This has been addressed with Federated Learning (FL), an emerging machine learning technology that enables distributed learning of a global prediction model without compromising privacy [Li *et al.*, 2020]. Given a group of users, FL approaches to HAR [Tu *et al.*, 2021] make use of local, user-specific supervision to update a global, high-quality activity predictor meant to be applicable to all users.

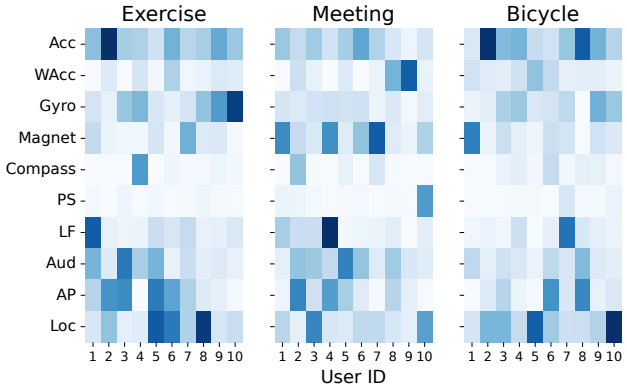

Figure 1: Importance of different features for 3 activities from 10 different individuals in ExtraSensory dataset. Saturation indicates higher relevance. The images indicate that the features important for recognizing any given activity strongly depend on the target user.

The above strategy, however, is not ideal in real-world settings, which are strongly characterized by cross-individual variability [Chen *et al.*, 2021; Zhang *et al.*, 2021b]:

Statistical Perspective. People are characterized by different habits, lifestyles and behavior patterns. From a HAR perspective, this means that the same activity may be performed very differently by different individuals, inducing a substantial cross-individual discrepancy in the conditional distribution of activities given sensor observations. In other words, the importance of specific sensors – and, by proxy, of latent features derived from them – is very individual-specific.

System Perspective. A major consequence of this fact is that, from the perspective of HAR *applications*, it is challenging to leverage statistical models learned on known users, for which annotated data is available, for predicting the activity of new users with their own activity characteristics.

Some FL-based approaches handle cross-individual diversity by learning user-specific models [Bettini *et al.*, 2021;

*Contact Author

Tu *et al.*, 2021]. Most importantly for our contribution, MetaHAR [Li *et al.*, 2021] trains a shared embedding network in a federated manner and then adapts the network with output layer to specific users via fine-tuning.

However, these approaches ignore the problem of *feature-level* discrepancy. In order to illustrate this problem, we trained a set of individual- and activity-specific random forest classifiers for three activities and ten randomly chosen individuals from the ExtraSensory dataset [Vaizman *et al.*, 2017], and visualized the contributions of different sensors in Fig. 1. Darker hues indicate higher relevance, computed as mean decrease in impurity. Consider the activity "exercise" (left image). It is clear that accelerometer data plays an important role in recognizing this activity for *User 2*, while predictions for *User 8* rely more heavily on GPS coordinates. The same pattern is clearly visible for all activities. This indicates that the sensors that contribute the most to recognizing certain activities strongly depend on the target individual. This observation is supported by recent studies on the diversity of human behavior in the social sciences, cf. [Zhang *et al.*, 2021b]. As a consequence, existing FL methods struggle when applied to individuals that were not observed at training time, as discussed in the Related Work section.

Prompted by these observations, we propose **FedMAT**, a **Fed**erated **M**ulti-task **AT**tention framework that extracts both shared and individual-specific features for multi-modal sensory feature fusion. FedMAT treats each individual as a learning task and utilizes a federated multi-task learning framework, considering that multi-task learning is naturally suited to the implementation of federated learning [Smith *et al.*, 2017]. FedMAT comprises a single shared network stored in a central node and multiple, individual-specific networks stored in distributed nodes. Specifically, the shared network is trained via federated learning across individuals to learn a set of global features. Then, for each individual an attention mask is applied to the shared network, such that each attention mask automatically learns the importance of the shared features for the different individuals. This way, FedMAT acquires a set of global features that are also shared and relevant for different subgroups of individuals, promoting generalization across individuals. We conduct an extensive empirical evaluation that shows FedMAT how outperforms several competitors when predicting activities of both known and new individuals ones, and performs especially well for challenging data sets collected under realistic, heterogeneous conditions. Our main contributions are as follows:

- We introduce FedMAT, a novel framework for cross-individual HAR that extracts and fuses individual-agnostic and individual-specific multimodal features in a federated multi-task learning manner.
- We propose a multi-task attention mechanism, which works as a mask for learning individual-specific features from the shared model while allowing for features to be shared among different individuals.
- We conduct extensive experiments on publicly available datasets. Results verify that FedMAT significantly outperforms baselines not only in generalizing to existing individuals but also in adapting to new individuals.

2 Related Work

2.1 Deep learning for HAR

Applications of Deep Learning (DL) to activity recognition are quite widespread [Wang *et al.*, 2019], owing to the ability of DL of achieving state-of-the-art performance without the need for explicit feature design [Zeng *et al.*, 2014]. A systematic evaluation of various feed-forward neural networks on activity recognition data shows that representation learning and time correlations are both critical to recognition performance [Hammerla *et al.*, 2016]. Building on this insight, models like DeepConvLSTM [Francisco and Daniel, 2016] and DeepSense [Yao *et al.*, 2017] leverage a hybrid architecture that combines CNNs and RNNs. AttenSense [Ma *et al.*, 2019] improves on these designs by implementing an attention module into a multimodal neural network in a way that is well suited for capturing both spatial and temporal correlations. Notably, the attention weights are identical across individuals. These approaches are not designed for handling cross-individual differences, and therefore may display degraded performance when applied in real-world HAR tasks, where the data is heterogeneous [Chen *et al.*, 2021].

2.2 Federated learning for HAR

Privacy-sensitive sensory information is a major challenge for traditional approaches to achieving high recognition accuracy while protecting users' privacy [Zhang *et al.*, 2021a]. Federated learning aims to train a centralized model using the data stored in multiple distributed nodes in a privacy-aware manner [Yang *et al.*, 2019]. Federated Averaging [McMahan *et al.*, 2017] combines local stochastic gradient descent (SGD) on client nodes with model averaging on the server-side, is able to reduce communication rounds between clients and server. Existing studies [Zhao *et al.*, 2018] have shown that federated learning performs well when clients hold non-IID data, and thus has some potential for addressing cross-individual diversity in HAR [Ouyang *et al.*, 2021]. However, although federated learning-based HAR approaches succeed in learning from different clients, pure FL does not model the similarity and discrepancy of the clients and thus fails to learn personalized models for all individuals. Multi-task learning (MTL) [Zhang and Yang, 2021] combines information from multiple, related learning tasks to improve prediction performance of all the tasks simultaneously, and represents a natural strategy for dealing with cross-individual differences, and exploiting cross-individual similarities, in HAR. MOCHA [Smith *et al.*, 2017] is proposed as a general federated multi-task learning framework and performs well for HAR task. However, it ignore the case of heterogeneous data distribution. Most closely related to our approach, MetaHAR [Li *et al.*, 2021] solves personalized HAR by treating each individual as a separate task and learns both shared and user-specific information. Compared with learning a one-size-fits-all model, MTL approaches can precisely capture relationships among non-IID data and are naturally well-suited for dealing with user heterogeneity in cross-individual HAR. However, existing works design the MTL model by taking a feedforward network and splitting the network at the *classification* layer, and therefore ignore discrepancy at the feature

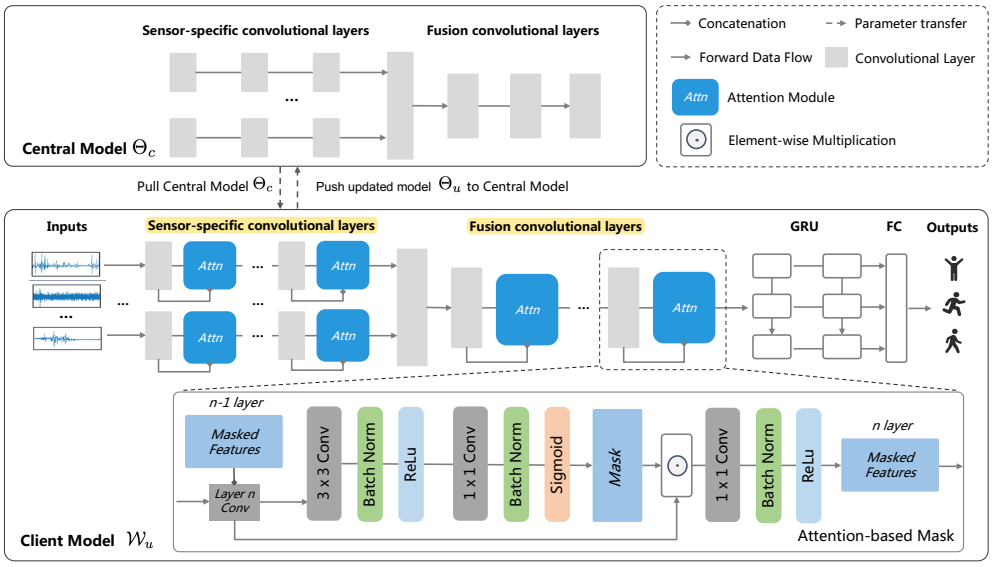

Figure 2: Architecture of FedMAT. Structures of the central model and one of the client models are visualized.

level.

3 Method

In this paper, we are concerned with cross-individual HAR. In the following, $\mathcal{U}$ indicates the set of individuals. For each such individual $u \in \mathcal{U}$, we have access to a corresponding data set $\mathcal{D}_u = \{(x_u^i, y_u^i)\}_{i=1}^{n_u}$, where $x_u^i \in \mathbb{R}^d$ are sensor readings and y_u^i are corresponding activity annotations.

3.1 Federated Multi-Task HAR

In our framework, we treat the HAR problem associated with each user as a separate learning task. In order to achieve strong generalization across individuals, we consider both heterogeneity and similarity between individuals by leveraging a Federated Multi-Task learning technique similar to MOCHA [Smith *et al.*, 2017]. The proposed architecture consists of a central model, with parameters Θ_c, and m decentralized models $\mathcal{W}_u, u \in \{1, 2, \cdots, m\}$ that learn individual-specific features. The overall goal is to acquire a HAR model that generalizes (i) across observed individuals, represented by $\mathcal{U}$, and (ii) to new individuals outside of $\mathcal{U}$. For now, let us focus on the first desideratum. This can be implemented by minimizing an appropriate loss over the observed individuals, as follows:

$$\min_{\Theta_c, \mathcal{W}_u} \sum_{u=1}^{m} \sum_{i=1}^{n_u} l_u(f_u(x_u^i; \Theta_c, \mathcal{W}_u), y_u^i). \quad (1)$$

Here, $l_u(\cdot, \cdot)$ indicates the loss function associated to user u, and $f_u(\cdot)$ a corresponding HAR predictor that depends on both the shared parameters Θ_c and the user-specific parameters $\mathcal{W}_u$. The architecture of the predictors is detailed next.

The embedding network processes the sensor observations x using a CNN-RNN architecture. As inputs for the neural networks, the training instances are partitioned by fixed-size sliding window into k time intervals of length L. This results in a data matrix of shape $d_s \times L$, where d_s is the dimension for each sensor s (e.g. x, y, and z axes for accelerometer). We then apply a Fourier transform to compute frequency-domain information, obtaining a final input tensor $\mathbf{X}_s$ of shape $d_s \times 2f \times k$, where f is the dimension of frequency-domain information. The set of tensors for each sensor, $\mathcal{X} = \{\mathbf{X}_s\}$, is finally the input of embedding network. The embedding network itself uses two sets of convolutional layers: the first set is applied to each sensor separately, the second one is applied to the concatenation of the individual sensor embeddings (see Fig. 2), so to fuse their representations and extract spatial dependencies between them. Within the two CNNs, we apply attention-based mask to extract individual-specific features, which will be introduced in Sec. 3.2. Then, Gate Recurrent Unit (GRU) layers are used to extract temporal relevance of the k CNN outputs. Finally, the embedding vectors output by the GRU layers are fed to a fully connected output layer that computes the probabilities for each category using a softmax activation.

Recall that our model is split into shared and individual-specific parts, which are stored separately. The central model Θ_c contains two CNNs that perform single sensor feature extraction and multiple sensor features fusion. As for the decentralized individual models $\mathcal{W}_u = \{a_u, h_u, c_u\}$, where a_u indicates attention-based mask modules for extracting individual-specific features, h_u indicates a GRU module for extracting temporal features, and c_u refers to output layer for classification. In this way, both individual-agnostic and individual-specific features can be extracted by the proposed framework. To optimize and update the model, the parameters are transferred between central server and distributed clients. Specifically, each individual with a local dataset $\mathcal{D}_u$ gets CNN models Θ_c from the central server, and introduce their data into CNNs masked by their local attention module to get their specific feature embeddings. Then, embedding vectors are introduced to GRU to get temporal features and

Algorithm 1 FedMAT.

Input: m individual-specific data sets $\{\mathcal{D}_u\}$, one per client.
Output: central model Θ_c, individual-specific models $\{\mathcal{W}_u\}$.

Server: Initialize central model $\Theta_c \leftarrow \Theta_0$
for $round = 1, 2, \ldots$ **do**
  for each $u \in \{1, 2, ..., m\}$ in parallel **do**
    Client u: Get central model Θ_c from the server.
    Client u: Train for n epochs using central model Θ_c together with local model $\mathcal{W}_u$, and get locally updated parameters Θ_u and $\mathcal{W}_u$.
    Client u: Push updated parameters Θ_u to server.
  end for
  Server: Update Θ_c according to Eq. 2
end for
return Θ_c and $\{\mathcal{W}_1, \ldots, \mathcal{W}_m\}$

finally get the loss via output layer. By performing n epochs of training locally in the clients, the parameters are separately updated to central server and decentralized nodes. The central server then averages the updates to update the shared model by averaging the models:

$$\Theta_c = \Theta_c + \lambda(\hat{\Theta} - \Theta_c), \tag{2}$$

$$\hat{\Theta} = \frac{1}{m}\sum_{u=1}^{m}\Theta_u. \tag{3}$$

3.2 Attention-based Mask

As mentioned above, the recognition of different individuals' activity relies on the different sensor readings. In order to precisely adapt the central model to unique individuals, we apply the attention-based mask to the feature representation layers, aiming at extracting individual-specific information. Therefore, we train multiple individual-specific attention networks. As such, the attention masks can be considered as feature selectors from the shared network, while the shared networks can learn a generalizing shared features across all individuals. Recall that our embedding network contains two types of CNN layers: (i) sensor-specific convolutional layers, and (ii) fusion convolutional layers. We apply attention-based mask module on both of the two types of convolutional layers, as shown in Fig. 2.

The detailed structure of the attention-based mask is shown in Fig. 2, consisting of multiple convolutional blocks for extracting task-specific features. Specifically, we refer the shared features in the l-th layer of the shared network as e^l, and the learned attention mask in this layer for individual u as e^l_u. The task-specific features $\hat{e}^l_u$ in this layer, are then computed by element-wise multiplication of the attention masks with the shared features:

$$\hat{e}^l_u = Mask^l_u \odot p^j. \tag{4}$$

For the first attention module in the convolutional layers, we takes as input only features in the shared network. As for subsequent attention mask in layer j, the input the concatenation of the shared features p^j, and the task-specific features from the previous layer $\hat{a}^{j-1}_i$:

$$Mask^l_u = h(g([p^l; f(\hat{e}^{(l-1)}_u)])). \tag{5}$$

Here, f, g, h are convolutional layers with batch normalization, following a non-linear activation ReLu in f, g or Sigmoid in h. Both f and g is composed with a $[3 \times 3]$ kernel, while h has a $[1 \times 1]$ kernel to match the channels between the concatenated features and the shared features. Then the attention mask $Mask^l_u \in [0, 1]$ is learned with back-propagation, which can operate as feature selectors from the shared features, while the shared network learns a generalized features across all individuals.

3.3 Individuals Adaptation

Finally, to address HAR task on new individuals, we treat the procedure as a meta-learning task. Specifically, we first meta-train the central feature representation networks Θ_c using the data from observed existing individuals as the procedure in Algorithm. 1. Then, for the new individual $\tilde{u}$, the shared networks Θ_c is then fine-tuned on the dataset of new individual $\mathcal{D}_{\tilde{u}}$ using *pairwise loss*:

$$l_{ij} = -\delta_{ij}\log(\sigma(D_{ij})) - (1 - \delta_{ij})\log(1 - \sigma(D_{ij})), \tag{6}$$

where $\sigma(\cdot)$ refers to the logistic sigmoid function, D_{ij} indicates the similarity between sample i and j, where we using cosine similarity, and $\delta_{ij} = 1$ if $y_i = y_j$, and 0 otherwise.

Then, the user-specific layers $\mathcal{W}_u$ together with the fine-tuned feature-representation network Θ_u trained with the *cross-entropy* loss. Back-propagation is applied to $\{\Theta_u, \mathcal{W}_u\}$ to minimize the classification loss.

4 Evaluation

In this section, we firstly introduce experimental setups and then describe the experimental results to answer empirically the following research questions:

Q1 Does FedMAT improve performance among existing individuals?

Q2 Does FedMAT help the adaptation to new individuals?

Q3 Does multi-task attention module learn heterogeneous features effectively?

Q4 How long does FedMAT take for adaptation?

4.1 Experimental Setup

We evaluate the proposed method on the following four wearable-sensor-based benchmark datasets:

- **HHAR [Stisen *et al.*, 2015]:** It contains $43,930,257$ accelerometer and gyroscope recordings collected from 9 individuals performing 6 activities. An important feature of this dataset is that users perform were asked to perform all activities while using 12 different devices.
- **PAMAP2 [Reiss and Stricker, 2012]:** It contains $3,850,505$ recordings from three inertial measurement units (IMUs) located on the hand, chest, and ankle. Each IMU hosts an accelerometer, gyroscope, magnetometer, thermometer, and heart rate sensor. The dataset was collected from 9 participants performing 12 main activities.

Model	HHAR		PAMAP2		ExtraSensory		SmartJLU	
	Accuracy	macro-F1	Accuracy	macro-F1	Accuracy	macro-F1	Accuracy	macro-F1
DeepSense	94.12	93.43	89.37	90.67	65.62	64.17	84.71	80.56
AttenSense	94.22	94.98	88.11	88.31	67.26	66.82	85.09	82.11
DeepSense-MTL	96.45	96.08	91.37	90.43	70.98	71.19	87.37	83.01
AttenSense-MTL	96.15	95.93	90.10	90.32	71.75	71.03	87.10	**84.32**
Meta-HAR	96.02	95.85	90.47	89.92	72.32	71.29	86.40	80.13
FedMAT-noSMask	96.17	96.01	91.89	91.73	71.36	70.43	87.82	83.79
FedMAT-noFMask	95.29	94.62	90.14	90.25	69.12	69.09	82.14	78.25
FedMAT	**96.88**	**96.81**	**92.61**	**91.84**	**75.72**	**75.03**	**89.78**	83.02

Table 1: Overall comparison results on generalizing with existing individuals (unit:%).

Model	HHAR		PAMAP2		ExtraSensory		SmartJLU	
	Accuracy	macro-F1	Accuracy	macro-F1	Accuracy	macro-F1	Accuracy	macro-F1
DeepSense	91.13	90.88	80.01	78.51	60.22	58.53	76.91	74.14
AttenSense	90.41	90.22	81.53	82.11	64.12	60.17	78.67	74.05
DeepSense-MTL	91.02	91.46	84.31	85.31	63.18	58.13	79.09	76.53
AttenSense-MTL	92.81	91.98	82.72	83.12	62.15	59.03	80.04	74.58
Meta-HAR	93.13	92.82	**86.91**	85.41	68.16	62.92	82.04	80.45
FedMAT-noSMask	95.77	95.56	83.89	82.73	71.36	68.43	85.33	83.59
FedMAT-noFMask	93.89	93.62	86.04	85.65	69.12	66.09	82.12	80.50
FedMAT	**95.83**	**95.81**	86.72	**85.94**	**73.83**	**69.97**	**86.74**	**84.55**

Table 2: Overall comparison results on adapting to the new individuals (unit:%).

- **ExtraSensory [Vaizman *et al.*, 2017]:** It contains over $300,000$ instances labeled with 51 types of human contexts and collected in a natural environment from 60 individuals. The records includes measurements from tri-axis sensors and information from smartphones and smartwatches.
- **SmartJLU:**[1] A similar dataset using the same tool and techniques as this one [Bison *et al.*, 2021] collected in China, which contains over $30,000$ instances labeled with daily activities collected from 50 individuals, over two weeks in a real-life scenario in which participants are required to use their smartphones naturally. During the data collection procedure, a smartphone app was used to carry out sensor recording (e.g. GPS, accelerometer) and administer periodic questionnaires about activity, location and social context. All students signed informed consent forms. The main features of this dataset are that it: (1) contains annotations for complex activities like "Housework"; (2) is collected in an unconstrained setup. In this experiment, the records are annotated with 23 different activities. The signals are obtained from smartphone sensors and include motion-reactive sensors (e.g., accelerometer), location, phone state, etc.

We evaluate FedMAT by comparing it with other models. Specifically, we compare both state-of-the-art models on HAR and two variants of FedMAT as follows:

- **DeepSense [Yao *et al.*, 2017]**: A deep learning model using CNN-RNN structure for sensor-based HAR.
- **AttenSense [Ma *et al.*, 2019]:** An attention-based multimodal neural network model for sensor-based HAR.
- **DeepSense-MTL:** Multi-task version of DeepSense.
- **AttenSense-MTL:** Multi-task version of AttenSense.
- **Meta-HAR [Li *et al.*, 2021]:** A federated representation learning framework, in which a signal embedding network is meta-learned in a federated manner.
- **FedMAT-noSMask:** FedMAT model removes the attention mask in sensor-specific convolutional layers.
- **FedMAT-noFMask:** FedMAT model removes the attention mask in fusion convolutional layers.

We implemented FedMAT using Python 3.6 and Pytorch 1.8. All experiments are carried out on a machine with 2 NVIDIA GeForce RTX 3090 GPUs. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\varepsilon = 10^{-8}$ is used to update all network parameters. For federated learning, we set $\lambda = 1.0$ and perform $n = 10$ epochs of local training at each update round.

We apply the settings of meta-learning by splitting all the users in a dataset into meta-train users, which participate in the meta-learning process, and meta-test users for testing the meta-learned model. For each user, we split the local dataset of each individual into a train set (80%) and a test set (20%).

[1]We open source SmartJLU dataset and source code on Github: https://github.com/Super-Shen/FedMAT.

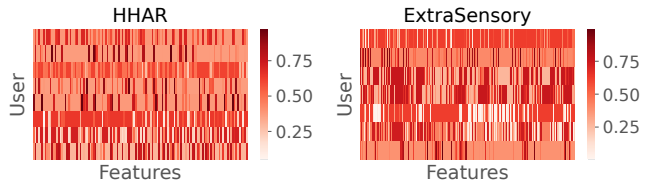

Figure 3: Visualization of attention-based mask on the HHAR and ExtraSensory datasets.

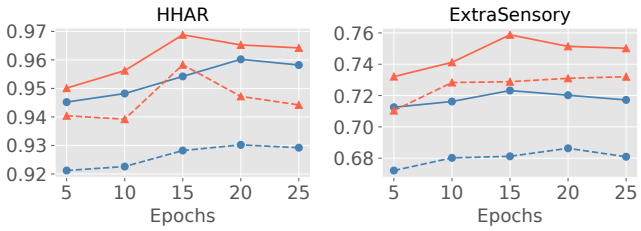

Figure 4: Evaluation of training epochs. Blue lines indicate performances of Meta-HAR and the red lines are for FedMAT. Dot lines refer to *macro-*F_1 and plain lines are for *accuracy*

To be specific, on the HHAR and PAMAP2 dataset, we randomly select one user as the meta-testing individual. For ExtraSensory dataset, seven activities are selected and nine individuals are randomly selected as meta-train set, while one individual is selected as meta-test set. For SmartJLU, nine individuals are randomly selected as meta-train, while two individuals for meta-testing.

Considering the labels are imbalanced, we use both *macro-*F_1 and *accuracy* as the performance metrics in the evaluation. We applied leave-one-individual-out validation. The reported performance is the average over all test individuals.

4.2 Experimental Results and Discussion

FedMAT improves performance for observed individuals. The results on generalizing among existing individuals are given in Table 1: (1) FedMAT outperforms generally baselines on four datasets, which means that learning multi-task attention modules in a federated learning manner can enhance the generalization ability among diverse individuals. (2) FedMAT is better than AttenSense, which shows that the attention weights on multimodal sensors can not be shared among all individuals and each individuals should be allocated with different attention because of the diverse behavior pattern. (3) FedMAT performs better than Meta-HAR, which shows that the heterogeneity should not only be considered in the classification layer, but also exits in the feature extraction procedure. (4) The fact that FedMAT works extremely well on ExtraSensory datasets than other competitors illustrates its robustness and generalization capability in real-world scenarios, where data distributions are largely heterogeneous.

FedMAT helps with adaptation to new individuals. As for the adaptation ability of FedMAT on new individuals, results are shown in Table 2: (1) FedMAT generally outperforms on four datasets, which means that it can handle datasets with high heterogeneity effectively and can be easily adapted to new individuals. (2) MTL models of DeepSense and AttenSense outperforms the single-task version, which indicates that MTL performs well on modelling features with heterogeneous data. (3) Meta-HAR works slightly better than FedMAT on the PAMAP2 dataset, since the signal distributions of PAMAP2 dataset are weakly heterogeneous. However, FedMAT can outperforms other models on more heterogeneous datasets, especially on real-world datasets.

FedMAT learns heterogeneous features effectively. We then evaluate the effect of multi-task attention module, we remove the attention modules in both sensor feature extraction layers and sensor features fusion layers. The results in Table 1 and Table 2 indicates that both of theses two attention modules contribute to the cross-individual HAR tasks. Generally, the attention module for sensor fusion layer has the higher impact. Moreover, to understand the role of the proposed attention modules, we visualize the attention masks $Mask \in [0, 1]$ of sensor fusion layer for each sensors across multiple individuals. As shown in Fig. 3, the different weights of various sensors are learned by our proposed approach. In particular, the attention masks have strong diversity across individuals, which validates the argument of the the motivating example in Section 1.

FedMAT adapts faster. We further compare FedMAT with the best contender, Meta-HAR, by increasing the number of client training epochs from 5 to 25 and check how accuracy changes. The results, illustrated in Fig. 4, verify that FedMAT consistently outperforms Meta-HAR on both controlled and real-world data sets. Moreover, FedMAT generally takes fewer epochs to achieve its best performance. This further stresses the effectiveness of our approach.

5 Conclusion

We introduced FedMAT, a novel federated learning framework for cross-individual sensor-based activity recognition that effectively addresses the heterogeneity in sensory feature distribution across different individuals. FedMAT works by extracting both shared and individual-specific features for multi-modal sensor fusion in the setting of FL. Our empirical evaluation shows that the proposed approach consistently outperforms several competitors on four different and challenging data sets.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (62077027), the Ministry of Science and Technology of the People's Republic of China(2018YFC2002500), the Jilin Province Development and Reform Commission, China (2019C053-1), the Education Department of Jilin Province, China (JJKH20200993K), the Department of Science and Technology of Jilin Province, China (20200801002GH), the European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us" (grant agreement No 823783), EU Horizon 2020 research and innovation programme TAILOR (grant number 952215).

References

[Bettini *et al.*, 2021] Claudio Bettini, Gabriele Civitarese, and Riccardo Presotto. Personalized semi-supervised federated learning for human activity recognition. *arXiv preprint arXiv:2104.08094*, 2021.

[Bison *et al.*, 2021] Ivano Bison, Fausto Giunchiglia, Mattia Zeni, Enrico Bignotti, Matteo Busso, and Ronald Chenu-Abente. Trento 2018 - an extended pilot on the daily routines of university students, 2021. University of Trento Technical Report - DataScientia dataset descriptors.

[Chen *et al.*, 2021] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(4):1–40, 2021.

[Francisco and Daniel, 2016] O. Francisco and R. Daniel. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[Hammerla *et al.*, 2016] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[Li *et al.*, 2021] Chenglin Li, Di Niu, Bei Jiang, Xiao Zuo, and Jianming Yang. Meta-har: Federated representation learning for human activity recognition. In *Proceedings of the Web Conference 2021*, pages 912–922, 2021.

[Ma *et al.*, 2019] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In *IJCAI*, pages 3109–3115, 2019.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Ouyang *et al.*, 2021] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 54–66, 2021.

[Reiss and Stricker, 2012] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[Smith *et al.*, 2017] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

[Stisen *et al.*, 2015] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pages 127–140, 2015.

[Straczkiewicz *et al.*, 2021] Marcin Straczkiewicz, Peter James, and Jukka-Pekka Onnela. A systematic review of smartphone-based human activity recognition methods for health research. *NPJ Digital Medicine*, 4(1):1–15, 2021.

[Tu *et al.*, 2021] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 15–28, 2021.

[Vaizman *et al.*, 2017] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing*, 16(4):62–74, 2017.

[Wang *et al.*, 2019] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[Yang *et al.*, 2019] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

[Yao *et al.*, 2017] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 351–360, 2017.

[Zeng *et al.*, 2014] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services*, pages 197–205. IEEE, 2014.

[Zhang and Yang, 2021] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[Zhang *et al.*, 2021a] Le Zhang, Wei Cui, Bing Li, Zhenghua Chen, Min Wu, and Teo Sin Gee. Privacy-preserving cross-environment human activity recognition. *IEEE Transactions on Cybernetics*, pages 1–11, 2021.

[Zhang *et al.*, 2021b] Wanyi Zhang, Qiang Shen, Stefano Teso, Bruno Lepri, Andrea Passerini, Ivano Bison, and Fausto Giunchiglia. Putting human behavior predictability in context. *EPJ Data Science*, 10(1):42, 2021.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.