HarMI: Human Activity Recognition Via Multi-Modality Incremental Learning

Xiao Zhang, Hongzheng Yu, Yang Yang, Jingjing Gu, Yujun Li, Fuzhen Zhuang, Dongxiao Yu, and Zhaochun Ren

***Abstract*—Nowadays, with the development of various kinds of sensors in smartphones or wearable devices, human activity recognition (HAR) has been widely researched and has numerous applications in healthcare, smart city, etc. Many techniques based on hand-crafted feature engineering or deep neural network have been proposed for sensor based HAR. However, these existing methods usually recognize activities offline, which means the whole data should be collected before training, occupying large-capacity storage space. Moreover, once the offline model training finished, the trained model can't recognize new activities unless retraining from the start, thus with a high cost of time and space. In this paper, we propose a multi-modality incremental learning model, called *HarMI*, with continuous learning ability. The proposed *HarMI* model can start training quickly with little storage space and easily learn new activities without storing previous training data. In detail, we first adopt attention mechanism to align heterogeneous sensor data with different frequencies. In addition, to overcome catastrophic forgetting in incremental learning, *HarMI* utilizes the elastic weight consolidation and canonical correlation analysis from a multi-modality perspective. Extensive experiments based on two public datasets demonstrate that *HarMI* can achieve a superior performance compared with several state-of-the-arts.**

***Index Terms*—Catastrophic forgetting, incremental learning, human activity recognition, mobile device, multi-modality.**

I. Introduction

HUMAN activity recognition (HAR) aims to detect human physical activities in real-world scenarios, which can allow intelligent systems to assist individuals with improvements of the quality of life in areas such as healthcare, smart cities, etc [16], [19], [24]. Human activity recognition (HAR) via smart sensing has drawn more and more researchers' interests both in academic and industrial areas in recent years [7], [32], [40]. Kinds of sensors (such as accelerometers, gyroscope, etc.) embedded in individuals' powerful smartphones or smart wearable devices are utilized to recognize human activities, which is widely used in many areas like medicinal services, business, security and so forth [40].

With the rapid development of deep learning [35], [36], [46], the existing works usually adopted kinds of deep neural network based on collected offline sensor data to recognize activities, for example, DeepSense [40], AttnSense [29]. However, due to the sensor data is sensitive to users' privacy, the HAR model tend to be transferred to edge devices and mobile devices locally rather than the remote servers in recent years [42]. Compared with the high-performance clusters, edge devices and mobile devices usually have very limited resources, for instance, limited storage, limited computing abilities, etc. When facing with resources limited environments, traditional methods encounter several challenges. In detail, traditional methods collected data from all kinds of sensors in all time intervals and built a general offline model to recognize activities, which consumed a lot of space to store huge training data. In addition, once the offline model training finished, it could be difficult for the trained model to recognize new activities unless retraining from the start, thus with a high cost of time and space. Therefore, to recognize activities incrementally along with the generation of sensor data is essential. Fig. 1 illustrates the difference between traditional offline HAR methods and incremental learning method. The users' activities could be a time series sequence: *driving, working, watching TV, etc.*, with data generated by multiple sensors. The incremental HAR models begin with an initial model, with new sensor data of a new activity generated, the model parameters could be updated to obtain abilities to recognize the new activity as well as the previous activities. Training an

Manuscript received January 23, 2021; revised May 15, 2021; accepted May 23, 2021. Date of publication June 1, 2021; date of current version March 7, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 62072279 and 62006118, in part by the Fundamental Research Funds of Shandong University, in part by CCF- Baidu Open Fund (CCF-BAIDU OF2020011), and in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20200460. *(Corresponding author: Yujun Li.)*

Xiao Zhang is with the School of Computer Science and Technology, Shandong University, Qingdao 266237, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, P.R. China (e-mail: xiaozhang@sdu.edu.cn).

Hongzheng Yu, Dongxiao Yu, and Zhaochun Ren are with the School of Computer Science and Technology, Shandong University, Qingdao 266237, China (e-mail: honnzhengyu@foxmail.com; dxyu@sdu.edu.cn; zhaochun.ren@sdu.edu.cn).

Yujun Li is with the School of Information Science and Engineering, Shandong University, Qingdao 266237, China (e-mail: liyujun@sdu.edu.cn).

Yang Yang is with the Nanjing University of Science and Technology, Nanjing 210014, China (e-mail: yangyfit@gmail.com).

Jingjing Gu is with the Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: gujingjing@nuaa.edu.cn).

Fuzhen Zhuang is with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China, and also with the Xiamen Data Intelligence Academy of ICT, CAS, China (e-mail: zhuangfuzhen@ict.ac.cn).

Digital Object Identifier 10.1109/JBHI.2021.3085602

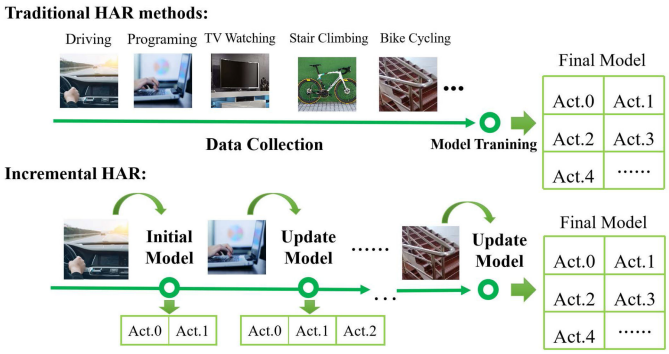


Fig. 1. An example to illustrate the difference between traditional offline HAR methods and incremental learning method.

HAR model incrementally means we don't need to keep the entire training data in memory, which is good for dealing with massive datasets or training in the resources limited embedding devices. In addition, the online nature also means that the model can quickly react to changes in the distribution of the newly generated sensor data.

Therefore, we consider to adopt incremental learning to recognize activities with continuous learning ability. The incremental learning methods try to retain knowledge learned from previous tasks during sequential learning process [13], [21], [25]. However, when applied incremental learning in multi-modality sensor data for activity recognition, we still encounter several challenges. (1) The sensors usually have different types and versions, which lead to the sampling frequencies of different sensors might be quite different. For instance, one HR monitor samples with 100HZ while another IMU sensor works in 33HZ, which means the amount of sampled data in one second is quite different. Therefore, measurements from multiple sensors are not aligned. (2) Catastrophic forgetting is common in incremental learning [10]. For example, an individual may sit in the office for 1 h, then walk 10 minutes to go home. We expect the model can discriminate sitting down in one time interval and walking in the next time interval. However, due to the fact that sensors generate measurements so fast, when changing to learn the model on walking, the model might forget the previous knowledge learned before in sitting. In addition, the sensing measurements are multi-modality, and how to overcome catastrophic forgetting in multi-modality scenario is challenging.

In this paper, we propose a multi-modality incremental learning model for HAR, called $HarMI$, to address the above challenges. First, we adopt an attention mechanism to weight each measurements in one time interval to align the data from multiple sensors. Next, we combine canonical correlation analysis (CCA) [18] and elastic weight consolidation (EWC) [22] to overcome catastrophic forgetting from a multi-modality perspective. In detail, we construct another attention layer for the fusion of each view's representation. In addition, we project each view's representation linearly and calculate the correlation coefficients among these projections. Based on the theory of CCA and method proposed by [20], we add the calculated correlation coefficient as regularization term of the loss function. Traditionally in incremental learning, EWC can consolidate parameters which have great influence on the loss of current task. Therefore we modify traditional EWC method to consolidate the parameters that have a great influence on the classification loss and the calculated correlations among modalities. Extensive experiments demonstrate the effectiveness of the proposed $HarMI$.

To sum up, our contributions can be summarized as follows.

- We address the HAR problem based on multiple sensors in incremental learning way. Compared with the traditional offline methods, to recognize activities incrementally can start training quickly and require little storage space, with high scalability when facing with new kind of activities. Solution to this problem is more similar with the real-world scenarios.
- We propose a multi-modality incremental learning model, named $HarMI$, to recognize activities based on heterogeneous sensing data from multiple sensors. Specifically, $HarMI$ aligns different sensor data based on attention mechanism. Besides, we attempt to solve the catastrophic forgetting problem of incremental learning from a multi-modality view. We combine CCA and EWC to consolidate model parameters that make contributions to reducing task loss and maintaining consistent outputs across each views.
- We evaluate our method on two public datasets, namely PAMAP2 and HHAR dataset, to verify the effectiveness and efficiency of $HarMI$.

The rest of the paper is organized as follows. Section 2 introduces the traditional human activity recognition methods, the related works on incremental learning and multi-modality learning. Section 3 formalizes the human activity recognition problem in incremental learning way. Section 4 proposes a multi-modality incremental learning model for HAR, namely $HarMI$, including aligning sensor data, multi-modality model with auxiliary output and consolidating parameters to overcome catastrophic forgetting. Section 5 reports experimental results, and demonstrates the performance of our proposed factor graph model. Section 6 concludes this paper.

II. Related Work

A. Human Activity Recognition

Multiple sensors based heterogeneous human activity recognition (HAR) has drawn many researchers' interests in the past few years. The existing methods for sensor based HAR were usually based on deep learning techniques, which can learn high level representations automatically. For example, [15] utilized ensembles of LSTM learners for activity recognition scenario to address the imbalanced datasets and the data quality problem. DeepSense [40] adopted CNN to learn the local interactions within each sensing modality and global interactions among different sensor inputs. In addition, RNN was utilized to learn the inter-interval relationships of the time series sensor data. AttnSense [29] introduced attention mechanism into CNN-RNN framework considering importance among different sensors and different time series measurements.

[5] proposed a multi-agent spatial-temporal attention model by considering the spatially-temporally varying salience of

features and the relations between activities and corresponding individual. [4] proposed an interpretable parallel recurrent model with convolution attentions for sensor based HAR on multi-modality scenario, in which CNN was used to extract the spatial relationships of features, and attention mechanism was utilized to extract the salient information of activities to address the problems of interperson variability and interclass similarity. However, Most of the previous HAR models usually built a general offline model to recognize activities while ignoring conducting HAR in a incremental learning way. The offline models should prepare the sensor data for training, usually occupying large-capacity storage space. In addition, when training new activities after the completion of training process, it must train start from scratch. Therefore, we propose a multi-modality incremental learning model for HAR based on kinds of sensor data.

B. Multi-Modality Learning

The multi-modality or multi-view data was very common in real-world scenarios. For instance, the audio signal and video signal can be regarded as two modalities in multimedia data. Similarly, different sensors were usually considered as different modalities in sensor based HAR. Traditionally, the single modality learning methods usually concatenated the multi-modality data into one feature vector, ignoring the specific characteristics of each modality. In contrast, multi-modality methods considered to improve the generalization performance by learning a function for each modality and optimizing all them jointly.

Traditionally, the multi-modality learning algorithms can be divided into three categories: co-training algorithms, co-regularization algorithms and margin-consistency algorithms [45]. In detail, the co-training algorithms were mainly proposed for semi-supervised learning [3], which utilized unlabeled data to train in two different modalities to maximize mutual consistency. Based on co-clustering framework and domain adaptation, IMAM [37] proposed a multi perspective adaptive model, which studied the domain adaptation of multi-modality data and realized the complementary transfer of cross domain knowledge in multiple subspace of features. IMAM used labeled data to learn a function for each modality and then predicted labels for unlabeled data. The co-regularization algorithms usually added regularization terms to the objective function to make sure that data from multiple modalities are consistent. Canonical correlation analysis (CCA) was a representative approach within co-regularization algorithms. As a regularization term, CCA seeks linear transformations for each modality, so the correlation among the transformed feature sets can be maximized in the common subspace while regularizing the self covariance of each transformed feature sets to be small enough [45]. Recently, with the fast development of deep learning techniques, lots of multi-modality methods based on deep neural networks have been proposed [2]. When applying CCA in deep neural networks, the traditional stochastic gradient descent can't work due to the fact that an optimization problem with constraint of covariance matrix needed to be solved. DCCA [2] utilized batch training to solve this problem but with very large calculation cost. This was because DCCA required to compute covariance matrices as well as inverse square roots for projection weights and even perform matrix singular value decomposition (SVD). Moreover, the performance of DCCA could be worse when the batch was too small. CorrReg [20] proposed a new method to calculate the correlation coefficients of the output projections of each modality to approximate the total correlation coefficients but with a much lower cost compared with DCCA [2]. Margin-consistency algorithms were proposed to utilize consistency of multi-modality data to regularize that the margins from two modalities should be the same or have the same posteriors [45]. For instance, MVMED [34] extends maximum entropy discrimination to multi-modality scenarios, enforcing the margins from two modalities to be identical.

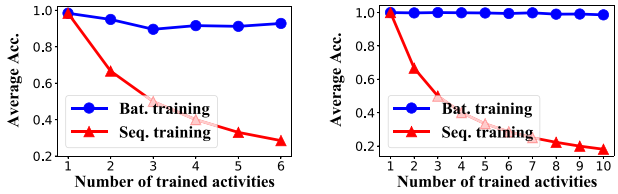
Some existing works considered recognizing activities based on multi-modality learning. For instance, Garcia *et al.* [11] utilized a multi-view stacking method to fuse the data from heterogeneous sensors for activity recognition. Kushwaha *et al.* [23] addressed the problem of silhouette-based HAR from multiple views, which used both contour-based pose features and uniform rotation local binary patterns for view invariant activity representation. However, most existing works usually built offline models, which consumed a lot of space to store huge training data. In addition, once the offline model training finished, it could be difficult for the trained model to recognize new activities unless retraining from the start, thus with a high cost of time and space. Therefore, to recognize activities in an incremental way along with the generation of sensor data is essential, which is the focus of work.

C. Incremental Learning

Incremental learning [27], also called continual learning [43] or lifelong learning [21], is a branch of online learning [17], usually learns the models through a sequence of tasks. The incremental learning model needs to retain knowledge of past tasks [22] when training on the current task. As mentioned in [6], incremental learning can be considered to continuously learn activities from newly coming data, however leading to catastrophic forgetting.

Catastrophic forgetting [1] [8], [14] is a common problem in incremental learning, which means the model can achieve the expected performance on the current task while performing worse on the previous tasks as if "forgetting" the learned knowledge from previous tasks. This is due to the fact that when training on a new task, model's parameters might be constantly updated according to the sensing data of current task. The average accuracy using batch training and sequential training under two public datasets for HAR is depicted in Fig. 2. As shown in the result, the average accuracy when training activities sequentially declines quickly, which demonstrates the catastrophic forgetting in sequential training. Traditionally, EWC [22] preserved knowledge of previous tasks by selectively slowing down learning on the weights important for those tasks, which is implemented with a form of quadratic regularization.

Incremental adaptive deep model (IADM) [38] was a method based on EWC regularization including attention mechanism.



(a) HHAR dataset (b) PAMAP2 dataset

Fig. 2. Average accuracy using batch training and sequential training under two public datasets for HAR. when training k-th activity, the batch training strategy (blue curve) use the data of the first activity to k-th activity, while the sequential training strategy (red curve) only uses the data of the k-th activity. The architecture of model is the same, using the attention layer to align sensor data, as discussed in Section IV-A (without auxiliary output layer and EWC). (a) HHAR dataset. (b) PAMAP2 dataset.

In detail, IADM constructed multiple classifiers with the deep layers of neural networks and used an attention layer to weight the outputs of those classifiers, which were also embedded in the calculated diagonal Fisher matrix. Compared with the traditional EWC, IADM has better scalability and sustainability. iCaRL [30] retained knowledge from previous tasks using knowledge distillation. However, iCaRL needed to keep a certain number of samples from previous tasks. With the increase of number of previous tasks, the number of samples in each task will become smaller, resulting in poor performance. PathNet [9] froze parameters of previous tasks and achieved knowledge sharing using a genetic algorithm to reuse some neurons, resulting in a fast consumption of model capacity. GEM [28] ensured that the updating of each parameter would not damage the performance of previous tasks and needed to preserve a certain amount of samples like iCaRL. GeppNet [12] used self-organizing mapping (SOM) in the input space to model the feature preferences of hidden layer neurons in the architecture. SOM network learned a representative vector for each task in the input space, and the representative vectors of similar tasks are as similar as possible. When learning a new task, the learning process was limited to a partition of the input space, so only part of the model parameters would be updated without affecting other parameters. Geepnet can still maintain high efficiency for high-dimensional data, but with the increase of the number of tasks, the performance maybe decrease obviously. DMA [26] was a dual memory architecture including a shallow neural network and a deep neural network. In detail, the shallow one can be trained quickly on the data of new tasks to meet the real-time requirements, while the deep network has larger parameter space achieving higher accuracy. However, DMA needed to store streaming data to train deep network, requiring extra large storage space.

However, the existing methods just considered to solve the catastrophic forgetting problem in single-modality scenario. Since the multiple sensors based HAR can usually be formalized as a multi-modality problem, utilizing the correlations among different modalities to address the catastrophic forgetting could be necessary. Therefore in this work, we combine EWC and canonical correlation analysis (CCA) to overcome catastrophic forgetting from a multi-modality perspective in sensor based HAR.

TABLE I
NOTATIONS USED IN THE PAPER

S	number of modalities
T	length of time window
x_i^j	the j-th data generated by i-th sensor in one time window
X_i	measurements generated by i-th sensor in one time window
XS_k^t	data generated in the t-th time window during the k-th activity
m_k	total number of time windows within the k-th activity
v_i	representation of the i-th modality
$\tilde{v}_i^l$	the l-th dimension of linear projection of the i-th modality's representation
$\tilde{\mu}_i^l$	average of $\tilde{v}_i^l$ in one mini-batch
σ_i^l	variance of $\tilde{v}_i^l$ in one mini-batch
λ_c	super-parameter on $Corr$ in training process
λ_e	super-parameter on EWC regularization term
$\theta_{Ac_k}^n$	the n-th parameter in the model after the training of the k-th activity.
$\mathcal{F}_{Ac_k}^n$	the n-th diagonal value of the diagonal Fisher matrix of k-th activity.

III. PROBLEM DEFINITION

Given a real-world scenario in which an individual might wear multiple sensors, the total number of sensors can be denoted as S. In every time window with T seconds ($T = 5\,s$ in default), the data generated by the i-th sensor is represented as $X_i = \{x_i^1, \ldots, x_i^j, \ldots, x_i^{n_i}\}$, in which x_i^j is a vector and n_i denotes the amount of the generated data. Due to the sampling frequency of each sensor is different, the amount of data generated in each time window also changes a lot. All data generated by all sensors in the time window can be represented by $XS = \{X_1, X_2, \ldots, X_S\}$. Assuming that individuals' activities are a sequence $Activity = \{Ac_1, Ac_2, \ldots, Ac_K\}$ with time, for example, $sitting, walking, running$, etc. A HAR model needs to be trained to recognize the total K activities in an incremental way. Within the k-th activity, the data generated by all sensors are denoted as $\{XS_k^t\}_{t=1}^{m_k}$, in which m_k is the total number of time windows with T seconds. The incremental HAR model is trained immediately after each activity finished, which means to using $\{XS_1^t\}_{t=1}^{m_1}$, $\{XS_2^t\}_{t=1}^{m_2}$, ..., $\{XS_K^t\}_{t=1}^{m_K}$ as training data respectively to train model in sequence. y_k represents label of the k-th activity, the goal of HAR is to learn a function that maps collected sensor data $\{XS_k^t\}_{t=1}^{m_k}$ to y_k in an incremental way:

$$f(\{XS_k^t\}_{t=1}^{m_k}) \rightarrow y_k \tag{1}$$

The notations used in the paper are shown in Table I.

IV. HARMI: MULTI-MODALITY INCREMENTAL LEARNING MODEL FOR HAR

In order to save storage space for training and learn new activities easily, we propose a multi-modality incremental model for human activity recognition, called $HarMI$, as shown in Fig. 3. The different sensors are considered as different views in $HarMI$. As mentioned before, the amount of sampled sensor data in one same time window is quite different. Besides, catastrophic forgetting is a common problem in incremental learning. To address the above challenges, the proposed $HarMI$ first introduces hierarchical attention layers for aligning the measurements from multiple sensors in one time window with T seconds.

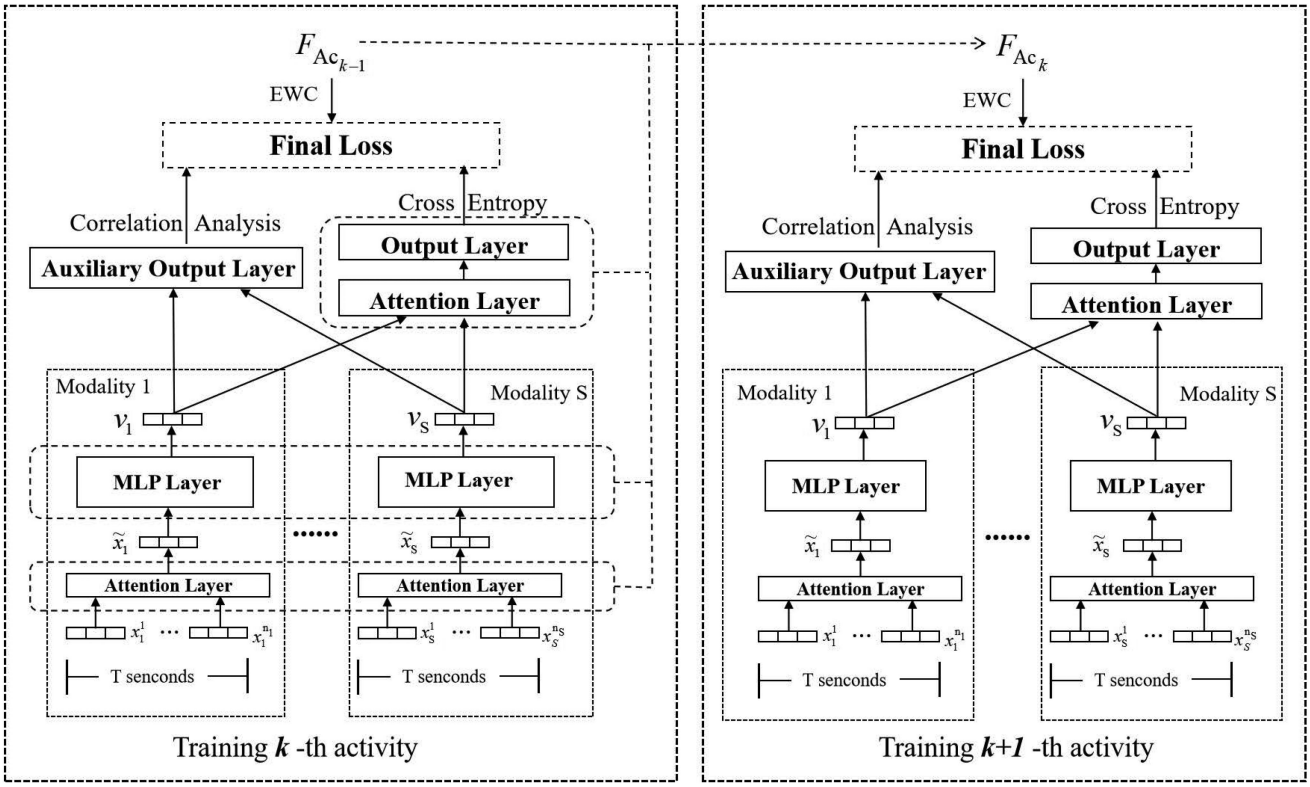


Fig. 3. The architecture of the proposed model *HarMI*. The solid line arrow denotes the the training process. The dotted arrows denote the parameter consolidation stage.

Then, in order to overcome the catastrophic forgetting problem, $HarMI$ combines CCA and EWC to consolidate parameters of previous activities from a multi-modality perspective. The details are shown as follows.

A. Aligning Sensor Data

Since each sensor might have different working frequency, the numbers of measurements generated by different sensors in one time interval could be quite different. In addition, different measurements generated by one same sensor in one same time interval could also contribute to the activity label differently. Therefore, we introduce the attention mechanism to align different sensor data as well as considering the imbalanced importance of measurements. The attention mechanism can assign the highest weight to the most important part of inputs, which has been applied in many applications successfully, such as image question answering [39], sequential recommendation system [41], or sensor based applications [44].

Therefore, during every time interval with T seconds, for the i-th sensor ($i \in \{1, 2, \ldots, S\}$), we utilize an attention layer to weight all the measurements and generate a fixed-length feature vector to achieve aligning of different sensors' data. As shown in Fig. 3, $HarMI$ adopts S attention layers to weight measurements for every sensor when training each activity. The details of the attention layer within each sensor can be formalized as follows:

$$\alpha_i^j = \frac{exp(\phi(w_i^1 x_i^j + b_i^1))}{\sum_{j=1}^{n_i} exp(\phi(w_i^1 x_i^j + b_i^1))} \tag{2}$$

$$\tilde{x}_i = \sum_{j=1}^{n_i} \alpha_i^j x_i^j \tag{3}$$

where $\phi(\cdot)$ denotes the activation function where we utilize the $tahn$ function to enhance non-linear capability; n_i denotes the total number of measurements generated by the i-th sensor; α_i^j refers to the importance of the j-th measurement; the intermediate representation of the i-th modality $\tilde{x_i}$ is the summation of the measurements weighted by the attention scores. and $\{w_i^1, b_i^1\}$ are parameters that need to be learned.

Then in order to improve the learning ability of the proposed model, we utilize a fully connected neural network with two hidden layers for each modality representation as follows.

$$v_i = MLP(\tilde{x}_i) \tag{4}$$

where $MLP(\cdot)$ represents the neural network with two hidden layers, v_i denotes the feature representation of the i-th modality, which is the transformation of the intermediate representation $\tilde{x}_i$. It is worth noting after transformation that $\{v_i | i \in 1, 2, \ldots, S\}$ have the same dimension.

B. Multi-Modality Model With Auxiliary Output

After calculating the representation of each modality within the training for the k-th activity, we use another attention layer to fuse the modality features. Different modalities could contribute differently to the prediction of the activities, and the attention layer can evaluate the importance of each modality. Then the final activity prediction can be obtained through an output layer

which is shown in Fig. 3. The details are shown as follows:

$$\beta_i = \frac{exp(\phi(w_i^2 v_i + b_i^2))}{\sum_{i=1}^{S} exp(\phi(w_i^2 v_i + b_i^2))} \tag{5}$$

where $\phi(\cdot)$ denotes the activation function as mentioned before; β_i is the importance measure of each modality, and $\{w_i^2, b_i^2\}$ are parameters which need to be learned.

$$y_k = w_{out}\left(\sum_{i=1}^{S} \beta_i v_i\right) + b_{out} \tag{6}$$

where y_k denotes the label of the k-th activity; $\{w_{out}, b_{out}\}$ are the parameters of the output layer that need to be learned.

In order to extract better advanced features, we want to keep the extracted modality representations consistent in a common space to maximize the correlations among each modalities [20]. Therefore, we construct an auxiliary output layer based on the method proposed in the literature [20], in which we project each modality representation linearly into one common space. By maximizing the correlation coefficient of these projections, we can make those outputs of modalities keep consistent, which is conducive to training process. Moreover, after the end of training process, it can also help overcome catastrophic forgetting, as we will discus in Section IV-C.

In detail, we take $\{v_i | i \in 1, 2, \ldots, S\}$ as the input of the CorrReg layer, which is used to calculate the correlations among different modalities. Each representation vector is linearly projected to a new one respectively, which is calculated as follows.

$$\tilde{v_i} = v_i w_i^* \tag{7}$$

in which $\tilde{v_i}$ denotes the linearly projected feature vector, and the w_i^* is the parameter corresponding to the i-th modality representation.

Then the correlations among each modality are estimated based on the linearly projected vectors $\{\tilde{v_i} | i \in 1, 2, \ldots, S\}$. The details are calculated as follows:

$$\tilde{\mu_i^l} = \frac{1}{N}\sum^{N} \tilde{v_i^l} \tag{8}$$

$$(\tilde{\sigma_i^l})^2 = \sum^{N} (\tilde{v_i^l} - \tilde{\mu_i^l})^2 \tag{9}$$

where the estimation of the correlation is calculated on each dimension and $\tilde{v_i^l}$ denotes the l-th dimension of $\tilde{v_i}$. Each $\tilde{v_i}$ has the same dimension with v_i. Given a mini-batch with N samples, the mean μ_i^l and variance σ_i^l on the l-th dimension are calculated.

Then the correlations among the modalities can be obtained:

$$Corr \approx \sum_{l} \frac{\sum^{N} \prod_{i=1}^{S} (\tilde{v_i^l} - \tilde{\mu_i^l})}{\sqrt{\prod_{i=1}^{S} (\tilde{\sigma_i^l})^2 + \epsilon}} \tag{10}$$

in which ϵ is utilized for the numerical stability. Similarly, the calculation of the correlation is also based on each dimension $\tilde{v_i^l}$ of the projected vectors.

Therefore, the predicted activity label and the correlation among modalities are calculated from the attention layer and the auxiliary output layer respectively. We specify the objective function as follows, in which the correlation added as a regularization term to improve the network training:

$$L = L_{obj} - \lambda_c Corr \tag{11}$$

in which L_{obj} denotes the classification loss and we use cross entropy in our work. L represents the final loss and λ_c is a superparameter.

C. Consolidating Parameters to Overcome Catastrophic Forgetting

As mentioned before, catastrophic forgetting is a common but tricky problem in incremental learning. This problem would also occur when training various kinds of activities sequentially. Therefore, in the $HarMI$ framework, we utilize a method based on EWC [22] combining with the calculated correlation among modalities in Section IV-B. However, the traditional EWC approach is utilized in single-modality scenario, ignoring the correlations among different modalities when facing with multi-modality sensor data. In detail, EWC method calculated the diagonal Fisher matrix based on the final loss at the end of each training iteration, in which each value in the matrix corresponds to a model parameter, reflecting its importance. Then in the next training iteration, a regularization term can be adopted to limit the change range of important parameters. When training the activities sequence $Activity = \{Ac_1, \ldots, Ac_K\}$ in an incremental way, we utilize the correlations among kinds of modalities to consolidate parameters. At the end of training iteration on the k-th activity Ac_k, we need to estimate the importance of each parameter about the current activity. Specifically, we pass all the training samples through our model to calculate the diagonal Fisher matrix $\mathcal{F}_{Ac_k}$. The n-th diagonal value of $\mathcal{F}_{Ac_k}$, represented as $\mathcal{F}_{Ac_k}^n$, can be calculated as follows:

$$\mathcal{F}_{Ac_k}^n = \sum_{t}^{m_k} \left(\frac{\partial L(XS_k^t)}{\partial \theta_{Ac_k}^n}\right)^2 + \mathcal{F}_{Ac_{k-1}}^n \tag{12}$$

where $L(XS_k^t)$ is the loss during evaluating the t-th sample of k-th activity, including L_{obj} and $Corr$ as shown in Eq. 11. $\theta_{Ac_k}^n$ is n-th parameter in the model after the training of the k-th activity. $\mathcal{F}_{Ac_{k-1}}^n$ is the saved diagonal Fisher matrix which are calculated after the training of the $k-1$-th activity. Specially, the $\mathcal{F}_{Ac_0}^n$ is a matrix with all zero, which is used to calculate $\mathcal{F}_{Ac_1}^n$ at the end of training process of the first activity.

In traditional EWC, the equation is as following:

$$L(\theta_{Ac_{k+1}}) = L_{obj}(\theta_{Ac_{k+1}}) + \lambda_e \sum_{n} \mathcal{F}_{Ac_k}^n (\theta_{Ac_{k+1}}^n - \theta_{Ac_k}^n)^2 \tag{13}$$

The more important the parameter for the old task, the larger the corresponding value in the Fisher matrix. Therefore, when using gradient descent to optimize the target loss $L(\theta_{Ac_{k+1}})$, these parameters that are important to the old task change little during the training process.

During the sequential training process, parameters in the hierarchical attention layers are continuously changed, leading to performance of the attention layer becomes worse when facing with the previous activities. In order to achieve better

performance, the representation of each modality needs to be consistent in the common space. Therefore in multi-modality scenarios, it is important to consolidate the parameters that can keep the representations of modalities consistent. So our proposed method calculates diagonal Fisher matrix based on two parts: the classification loss and a regularization term $Corr$ which denotes the estimation of correlations among each modality. The parameters relate with the classification loss and view output consistency can be consolidated. In the next training iteration, the proposed $HarMI$ will retain the above parameters preferentially, which can overcome the catastrophic forgetting problem. When training on the $k+1$-th activity Ac_{k+1}, we use another regularization term to consolidate the important parameters of the previous k activities. The loss function can be calculated as:

$$L(\theta_{Ac_{k+1}}) = L_{obj}(\theta_{Ac_{k+1}}) - \lambda_c Corr + \lambda_e \sum_n \mathcal{F}^n_{Ac_k}(\theta^n_{Ac_{k+1}} - \theta^n_{Ac_k})^2 \quad (14)$$

in which $\theta^n_{Ac_{k+1}}$ represents the n-th parameter of current model, $\theta^n_{Ac_k}$ denotes the corresponding saved parameter of activity Ac_k, $\mathcal{F}^n_{Ac_k}$ represents the corresponding value of $\theta^n_{Ac_k}$ in the diagnosis Fisher matrix of activity Ac_k. λ_e is a super-parameter deciding importance the previous k activities comparing with the new one. Then, in the same way, we calculate diagnosis Fisher matrix $\mathcal{F}_{Ac_{k+1}}$ as shown in Eq.12 and update saved parameters of previous activities to current model parameters, which would be repeated during the incremental training process until all activities are trained.

V. Experiment

In this section, we evaluate the proposed $HarMI$ on two benchmark datasets. We first detail the experimental setup, dataset description and the adopted baselines, respectively. Then we demonstrate the effectiveness of our multi-modality incremental learning model.

A. Experimental Setup

We train the model in sequential style $Activity = \{Ac_1, \ldots, Ac_K\}$ one by one varying with time. After finishing the training of each activity, the model will not store the training data again. We conduct the experiments in two scenarios. One is without repetitive activities, which means the kinds of activities in the sequence are all different. The other one is with repetitive activities, in which each kind of activity appear twice and then the activity sequence is permuted randomly.

In detail, we use measurements generated by multiple sensors within one time window ($T = 5$ $seconds$ in default) as input for activity recognition. We use mini-batch to train the model. Since EWC constraints are used to solve the catastrophic forgetting problem, SGD is chosen as the optimizer. In order to find the model's super-parameters, we use grid search to find the super-parameter of EWC constraint regularization, learning rate and momentum. We try [100, 500, 2500, 5000, 8000, 10 000, 15 000, 20 000] super-parameters respectively for the EWC regularization, [0.0001, 0.000 001] for learning rate and [0, 0.9, 0.99] for momentum. It is worth noting that we need to find the appropriate three super-parameters again when training each activities. To sum up, for learning each activitiy, we need $8 \times 2 \times 3 = 48$ search process. We adopt the *accuracy*, *precision*, *recall*, and *F1-score* as the evaluation metrics.

TABLE II
NUMBER OF USED PARAMETERS OF ALL MODELS IN TWO DATASETS

	HarMI	Mm-EWC	Sm-EWC	DeepSense
HHAR	900,768	900,768	900,829	900,225
PAMAP2	1,200,244	1,200,244	1,200,253	1,201,496
	LWF	iCaRL	IADM	GeppNet
HHAR	900,768	900,768	900,663	949,943
PAMAP2	1,200,244	1,200,244	1,201,843	1,219,791
	HarMI-GRU			
HHAR	901,020			
PAMAP2	1,201,101			

Number of parameters: In general, models with more parameters will retain more knowledge about previous tasks and generally perform better. To make a fair compare on our model and baselines, we keep all model parameters within the same range on both datasets. Table II show number of parameters for each model in two datasets.

Early stop: The number of epochs for training each activities is set as 500. Due to the strict constraints of the consolidated parameters, it would be difficult for the model to learn knowledge on current activity. Therefore, model's parameters are recorded only if accuracy on the current activity is greater than a threshold (we set as 0.4 in the experiment). Since the accuracy on previous activities is very sensitive to the parameters updating. During training process on current activity, if the accuracy is very large, the performance on previous activities usually decline sharply. Therefore, in each iteration, if the accuracy on current activity is greater than a threshold (we set as 0.95 in the experiment), training process will be early stopped.

B. Dataset Description

We evaluate our model on two public datasets for HAR, which are shown as follows.

- **PAMAP2** [31] The Physical Activity Monitoring dataset contains data of 12 different physical activities wearing 3 inertial measurement units (IMU) and a heart rate monitor (HR). Three IMU monitors (worn on hand, chest and ankle) work with 100HZ while HR monitors are with 33HZ. Each inertial measurement unit (IMU) are recognized as one modality which contains 10-dimension measurements totally (3D-accelerometer data, 3D-gyroscope data, 3D-magnetometer data and 1D-temperature data).
- **HHAR** [33] The Heterogeneity Human Activity Recognition dataset contains data on 6 different activities from accelerometer and gyroscope from 12 devices (8 smartphones and 4 smartwatches). Smartphones and smartwatches all generate 6-dimension measurements including 3D-accelerometer data and 3D-gyroscope data. The measurements from accelerometer and gyroscope are recognised as two modalities respectively.

C. Compared Baselines

We evaluate the performance of the proposed multi-view incremental learning model with the baselines shown as follows.

- **DeepSense [40]:** It was a framework with an architecture using both CNN and RNN to capture relations among different sensors and time-series data on HAR tasks.
- **Single-modality EWC:** We simply concatenate aligned measurements of each modality and construct a single view model to sequentially train HAR models. EWC constraint is used in this model.
- **Multi-modality EWC:** We use multi-view neural network but using EWC loss to sequentially train HAR models. This is exactly the same structure as $HarMI$ except that CCA constraints are not added.
- **HarMI-RNN:** We use the gated recurrent unit recurrent neural networks (GRU) to extract advanced features from time-series sensing data rather than the attention mechanism used in *HarMI*. GRU is an variant of the LSTM network but has a simpler structure. The rest architecture of model is exactly the same as $HarMI$.
- **LWF [27]:** A method can be seen as a combination of distillation networks and fine-tuning. It dynamically adds new output layers to learn new tasks. First of all, network parameters except those in the new output layer are frozen and fine-tuning is used to train the new output layer parameters until convergence. Then the old network is used as teacher network, adding distillation loss into loss function and only using the data of new tasks to train the whole network until convergence. The architecture of model is same with *HarMI*
- **iCaRL [30]:** It is an incremental learning method based on pseudo-rehearse way, which needs to retain the previous training data. It uses distillation network to learn new tasks in training state, and uses metric learning method to classify. We preserve 77 350 records of PAMAP2 dataset and 41 000 records of HHAR dataset for models' training. SGD optimizer with learning rate 0.1 are used in training process. The architecture of model is same with *HarMI*
- **IADM [38]:** It is an incremental adaptive deep model (IADM) which introduced attention mechanism into EWC to overcome Catastrophic Forgetting in incremental learning. It contains multiple classifiers, and the prediction is the weighted average of predictions from multiple classifiers using attention mechanism. While computing the fisher matrix, it would incorporate the learned attention weight into the corresponding model parameters in fisher matrix.
- **GeppNet [12]:** It is a biologically inspired architecture for incremental learning. The structure is much simple, which projects the input vector into the SOM network then predict a label using the linear regression. Considering that the number of parameters of each method on the same dataset are kept as the same as possible, we use 17×17 SOM network on HHAR dataset (949 943 parameters) and 9×9 SOM network on PAMAP2 dataset (1 219 731 parameters) respectively.

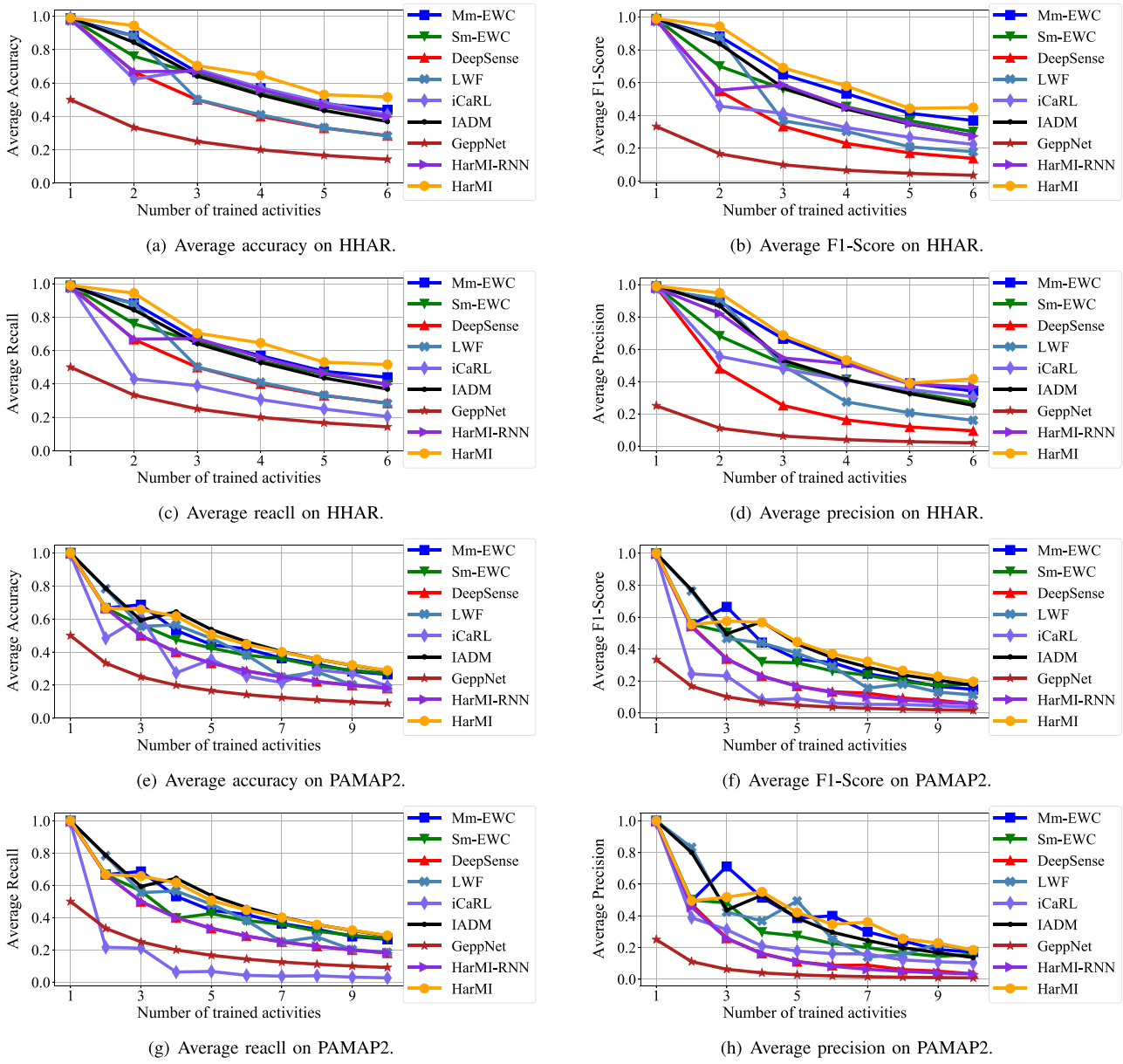
D. Numerical Results

Experimental scenarios without repetitive activities: Fig. 4(a)–Fig. 4(d) are average results on four metrics on HHAR dataset, while Fig. 4(e)–Fig. 4(h) are results on PAMAP2 dataset.

- With the increase of number of trained activities, the performance of all methods tend to become worse because of catastrophic forgetting. However, the proposed multi-modality incremental learning model $HarMI$ decline more slowly on four metrics compared with other baselines, which can achieve the best performance.
- Since $DeepSense$ is not designed for the incremental learning scenario, the performance of $DeepSense$ declines sharply when facing with the training of sequential activities. As shown in the result, the proposed $HarMI$ can achieve 23.13% accuracy, 31.04% F1-score, 23.13% recall, 32.16% precision improvement compared with $DeepSense$ when finishing the training of all activities.
- The performance of *Multi-modality EWC* is close to that of our proposed $HarMI$. This is due to the fact that *Multi-modality EWC* method just lacks the auxiliary output layer, in other words, the correlation regularization term comparing with $HarMI$. To consolidate the parameters that can keep modalities consistent could definitely help improve the performance when facing with multi-modality sensor data in sequential training.
- On $PAMAP2$ dataset shown in Fig. 4(e)–Fig. 4(h), from one trained activity to two trained activities, four metrics all decline sharply. This is because that the first activity is $sitting$, which is very similar with the second activity $standing$ that makes $HarMI$ difficult to classify the previous activities.
- The performance of *HarMI-RNN* is worse than *HarMI*, which shows that the attention mechanism is more suitable to handle time-series sensing data in incremental learning than the RNN. Attention has a simpler structure which can converge easier and faster.
- Fig. 6 depicts the confusion matrix of $HarMI$ after sequentially training all activities on $HHAR$ dataset. We can see that $HarMI$ can maintain the knowledge of previous activities and alleviate forgetting to a certain extent. The accuracy of activities like *stand*, *sit*, *stairs up* and *bike* are higher than 50%. However, the forgetting problem cannot be completely prevent during the incremental learning process. The activities like *walk* and *stair down* is quite difficult for $HarMI$ to classify.

Experimental scenarios with repetitive activities: Fig. 5(a)–Fig. 5(d) are average results on four metrics on HHAR dataset, while Fig. 5(e)–Fig. 5(h) are results on PAMAP2 dataset.

- The proposed $HarMI$ can achieve the best performance compared with other baselines on four metrics.
- When facing with the repeated activities, the performance of $HarMI$ can keep in steady level. For example on $HHAR$ dataset shown in Fig. 5(a)–Fig. 5(d), from three trained activity to four trained activities, four metrics



(a) Average accuracy on HHAR. (b) Average F1-Score on HHAR.

(c) Average reacll on HHAR. (d) Average precision on HHAR.

(e) Average accuracy on PAMAP2. (f) Average F1-Score on PAMAP2.

(g) Average reacll on PAMAP2. (h) Average precision on PAMAP2.

Fig. 4. Experimental results on two public datasets: HHAR and PAMAP2 without repeated activities ($T = 5\,s$ in default). (a) Average accuracy on HHAR. (b) Average F1-Score on HHAR. (c) Average reacll on HHAR. (d) Average precision on HHAR. (e) Average accuracy on PAMAP2. (f) Average F1-Score on PAMAP2. (g) Average reacll on PAMAP2. (h) Average precision on PAMAP2.

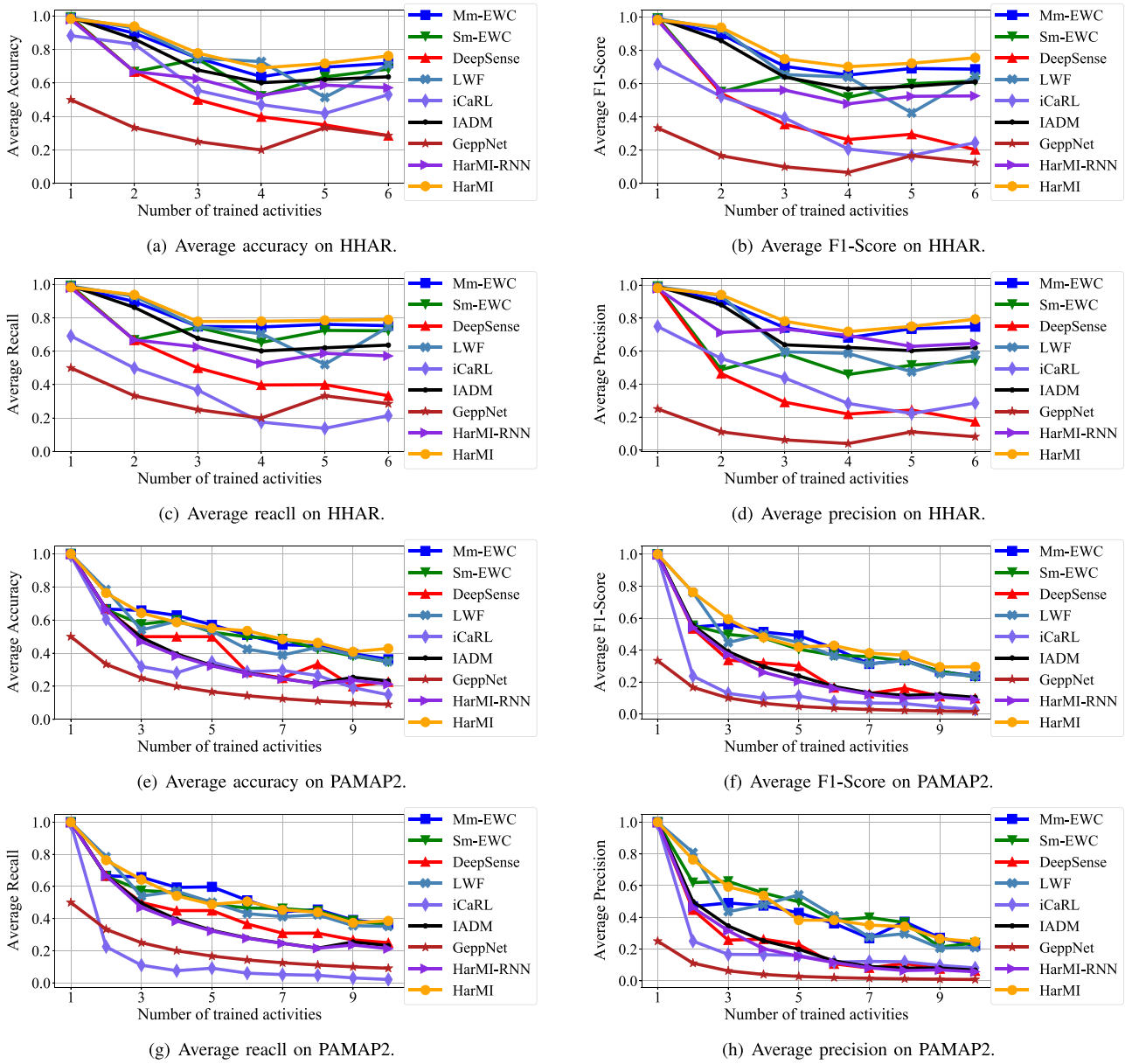
all decline because of catastrophic forgetting. However, from four trained activity to five trained activities, the performance become better. This is due to the fact that the third activity and the fifth activity all denote $walk$. When training on repeated activities, the $HarMI$ has consolidate the important parameters for $walk$.

E. Parameter Analysis

To investigate the influence of different parameters that might influence the performance of $HarMI$, we conduct experiments on following parameters:

Length of time window T: The influence of length of time window T on the performance is shown in Table III, in which T varies from 3 s to 7 s. We can see that the proposed $HarMI$ can outperform all the other baselines on both datasets under different length of time window. With the increment of T, both accuracy and F1-score of $HarMI$ tend to become larger as shown in the result. It is because that the larger T denotes that more sensor data in this time window can be used for model training.

Attention weight: The comparison of the weights for the attention layer on the modality fusion between $HarMI$ and *Multi-modality EWC* is depicted in Fig. 7. As mentioned before, the *Multi-modality EWC* method just lacks the auxiliary output layer compared with $HarMI$. As shown in the result, on $PAMAP2$ dataset, the attention weights of different views in $HarMI$ vary a lot. During the sequential training, *modality*



(a) Average accuracy on HHAR. (b) Average F1-Score on HHAR.

(c) Average reacll on HHAR. (d) Average precision on HHAR.

(e) Average accuracy on PAMAP2. (f) Average F1-Score on PAMAP2.

(g) Average reacll on PAMAP2. (h) Average precision on PAMAP2.

Fig. 5. Experimental results on two datasets: HHAR and PAMAP2 with repeated activities ($T = 5\,s$ in default). (a) Average accuracy on HHAR. (b) Average F1-Score on HHAR. (c) Average reacll on HHAR. (d) Average precision on HHAR. (e) Average accuracy on PAMAP2. (f) Average F1-Score on PAMAP2. (g) Average reacll on PAMAP2. (h) Average precision on PAMAP2.

0, namely the *IMU sensor on the hand* modality, is the most important, followed by the *IMU sensor on the chest* modality and the *IMU sensor on the ankle* modality. However, *Multi-modality EWC* method gives each modality equal weights, which leads to the relative worse performance.

F. Comparison With Traditional Batch Training Method

To train the HAR model in an incremental way is necessary, and particularly suitable in resources limited environments, such as wearable devices, edge devices. When training the k-th activity, $HarMI$ only uses the training data of current activity while the batch training methods need to use all the training data of the previous k activities. Therefore, we compare the proposed $HarMI$ with traditional batch training methods (DeepSense is utilized in this work) considering both training time cost and storage cost. The results are described as follows.

1) Training Time Cost: Since individuals' activities are a sequence $Activity = \{Ac_1, Ac_2, \ldots, Ac_K\}$ with time, for example, *sitting, walking, running*, along with the sensor data. With the increase of the number of activities, the training time cost denotes the spent time from the start of training process to the convergence on current activity. SGD optimizer is utilized to update parameters. For $HarMI$, we set learning rate as 0.001, and λ_e as 5000 on two datasets mentioned before. And we set the initial learning rate of batch training model as 0.1 for comparison. In addition, we adopt dropout in the proposed $HarMI$ framework, in which the rate is set as 0.2 in the first

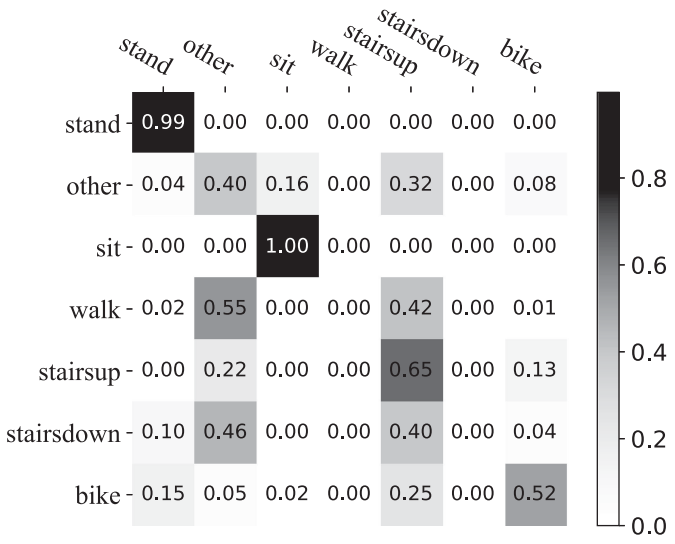


Fig. 6. Confusion matrix of $HarMI$ at the end of sequential training for all activities in HHAR dataset.

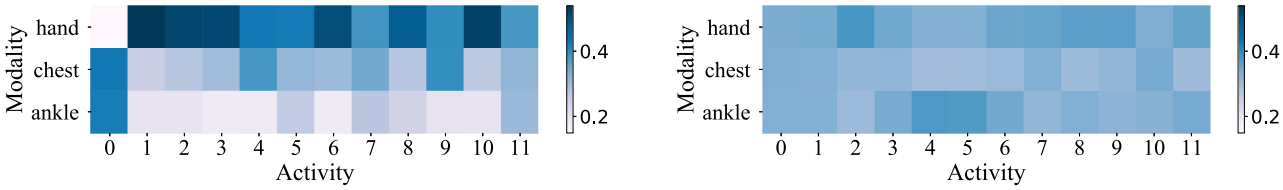
TABLE III
EXPERIMENTAL RESULTS WITH DIFFERENT LENGTHS OF TIME WINDOWS ($T = 3\,s, 5\,s, 7\,s$) ON THE TWO PUBLIC DATASET: HHAR AND PAMAP2 WITHOUT REPEATED ACTIVITIES

DATASET		HHAR			PAMAP2		
T		3S	5S	7S	3S	5S	7S
HarMI	Acc.	**49.84**	**51.61**	**50.33**	28.81	**28.92**	**33.19**
	F1	**43.67**	**44.89**	**45.84**	**18.15**	**19.48**	**22.16**
	Pre.	**42.08**	**41.66**	**48.36**	13.89	**18.41**	**18.74**
	Rec.	**49.84**	**51.61**	**50.33**	**28.81**	**28.92**	**33.19**
Mm-EWC	Acc.	45.35	43.97	46.27	26.96	26.51	25.61
	F1	36.04	36.97	37.49	16.40	14.66	14.64
	Pre.	36.69	34.11	36.19	**16.04**	17.15	12.28
	Rec.	45.35	43.97	46.27	26.96	26.51	25.61
Sm-EWC	Acc.	33.67	39.54	38.71	27.54	26.98	29.12
	F1	22.48	30.08	29.84	15.71	18.07	15.54
	Pre.	21.85	26.91	27.53	11.38	15.61	10.72
	Rec.	33.67	39.54	38.71	27.54	26.98	29.12
DeepSense	Acc.	28.48	28.48	28.48	18.07	18.07	18.07
	F1	13.86	13.86	13.86	5.65	5.65	5.65
	Pre.	9.49	9.49	9.49	3.35	3.35	3.35
	Rec.	28.48	28.48	28.48	18.07	18.07	18.07
LWF	Acc.	28.25	28.26	48.90	21.56	18.44	19.21
	F1	17.90	17.96	37.25	12.67	11.21	11.30
	Pre.	15.88	15.97	31.02	10.40	17.88	10.70
	Rec.	28.25	28.26	48.90	21.56	18.44	19.21
iCaRL	Acc.	32.18	41.02	41.45	**39.02**	19.42	25.29
	F1	8.40	22.44	20.67	7.93	3.48	6.49
	Pre.	17.51	30.61	28.36	13.89	10.24	12.98
	Rec.	6.29	20.48	18.65	5.97	2.66	5.72
IADM	Acc.	28.27	36.92	31.35	24.93	28.67	22.47
	F1	17.76	27.68	20.39	11.64	17.18	9.78
	Pre.	15.57	25.25	16.41	7.87	13.67	6.32
	Rec.	28.27	36.92	31.35	24.93	28.67	22.47
GeppNet	Acc.	14.28	14.28	14.28	9.09	9.09	9.09
	F1	3.57	3.57	3.57	1.51	1.51	1.51
	Pre.	2.04	2.04	2.04	0.82	0.82	0.82
	Rec.	14.28	14.28	14.28	9.09	9.09	9.09
HarMI-RNN	Acc.	43.39	39.82	35.13	18.09	18.18	23.76
	F1	34.03	27.51	24.77	8.07	5.61	12.06
	Pre.	31.35	36.59	20.71	5.80	3.32	8.44
	Rec.	43.39	39.82	35.13	18.09	18.18	23.76

layer, and 0.5 in the deep layers. It is worth noting that dropout can make the optimizer randomly select part of parameters in the updating phase, which is used to prevent forgetting of previous activities. However, the adoption of dropout can also affect the convergence rate and make the training time become longer. Therefore, we conducted two groups of experiments: one with dropout and the other without.

The training time cost with the increase of number of activities on two datasets are depicted in Fig. 8 and Fig. 9. In detail, Fig. 8 and Fig. 9 depict the average time cost per epoch and total time cost on two public datasets. The total time cost denotes the spent time until the convergence of models when training the k-th activity, which is the product of the average time cost per epoch and the number of trained epochs. As shown in the results, regarding batch training method, we can see that both average time spent per epoch and total time cost increase with the number of activities, while $HarMI$ keeps relatively stable. This is due to the fact that when training a new activity every time, the batch training method needs to use all the data from all previous activities while the training of $HarMI$ is just based on the data of current activity. Therefore, the increasing amount of data leads to higher time cost. Within $HHAR$ dataset, after finishing training the last activity, the average time cost per epoch with dropout of batch training method is about 2.1 times than that of $HarMI$ (1.7 times without dropout), and the total time cost with dropout of batch training is about 57 times than that of $HarMI$ (59 times without dropout). The time cost difference is even larger in $PAMAP2$ dataset. The average time cost of batch training method is about 4 times (with dropout), and 4.5 times (without dropout) comparing with $HarMI$. Furthermore, the total time cost of batch training is 101 times (with dropout), and 66 times (without dropout) than $HarMI$.

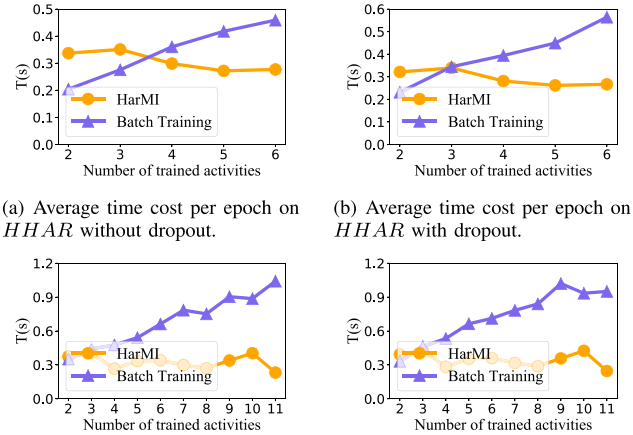
2) Storage Cost: Storage is another kind of important resource within IoT devices. Therefore, we also measure the storage cost on $HarMI$ and the traditional batch training method. We consider the number of stored data samples as the storage cost during the training of the k-th activity. The results are shown in Fig. 10. As for batch training methods, after learning a new activity, the training data would be reserved for another round of training. Therefore, with the increase of the number of activities, the storage cost becomes larger. As shown in the result, on $HHAR$ dataset, more than 2 million training samples need to be stored after learning 7 activities, while storing more than 5 million training samples after learning 12 activities in $PAMAP2$ dataset. Therefore, for devices with limited storage capacities, e.g. mobile devices and kinds of sensors, with the increase of the number of activities, the batch training methods might not work to train due to the lack of storage space. In comparison, the proposed $HarMI$ only needs the training data of the current activity during the training stage. And once finishing the training of current activity, the training data can be dropped. As can be seen from the figure, with the increase of the number of activities, the storage cost of $HarMI$ keeps in quite low level. On $HHAR$ dataset, the maximum storage cost is 420 thousand training samples, while it is 630 thousand on $PAMAP2$ dataset. Therefore, for the devices with limited storage, even if the training data of all activities cannot be



(a) Attention weight matrix of *HarMI*.

(b) Attention weight matrix of *Mm-EWC*.

Fig. 7. Comparison of the weights of attention layer on modality fusion between *HarMI* and *Mm-EWC* on PAMAP2 dataset. (a) Attention weight matrix of *HarMI*. (b) Attention weight matrix of *Mm-EWC*.



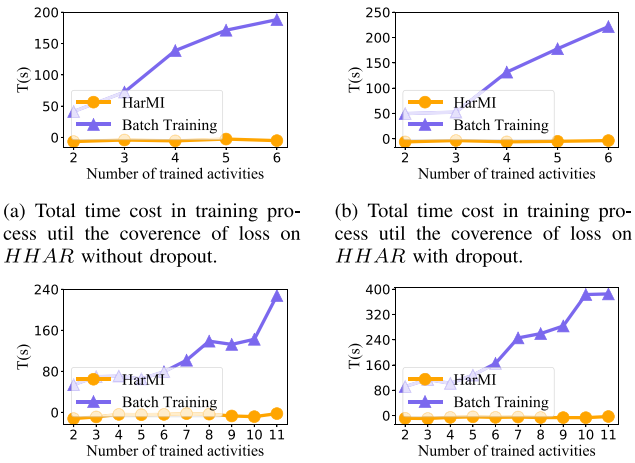
(a) Average time cost per epoch on $HHAR$ without dropout.

(b) Average time cost per epoch on $HHAR$ with dropout.

(c) Average time cost per epoch on $PAMAP2$ without dropout.

(d) Average time cost per epoch on $PAMAP2$ with dropout.

Fig. 8. Average time cost per epoch with the increase of number of activities on two datasets: $HHAR$ and $PAMAP2$ during training. (a) Average time cost per epoch on $HHAR$ without dropout. (b) Average time cost per epoch on $HHAR$ with dropout. (c) Average time cost per epoch on $PAMAP2$ without dropout. (d) Average time cost per epoch on $PAMAP2$ with dropout.



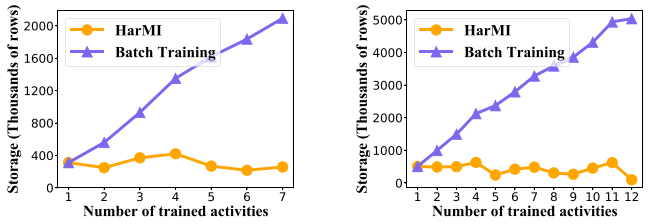
(a) Total time cost in training process util the coverence of loss on $HHAR$ without dropout.

(b) Total time cost in training process util the coverence of loss on $HHAR$ with dropout.

(c) Total time cost in training process util the coverence of loss on $PAMAP2$ without dropout.

(d) Total time cost in training process util the coverence of loss on $PAMAP2$ with dropout.

Fig. 9. Total time cost with the increase of number of activities on two datasets: $HHAR$ and $PAMAP2$. (a) Total time cost in training process util the coverence of loss on $HHAR$ without dropout. (b) Total time cost in training process util the coverence of loss on $HHAR$ with dropout. (c) Total time cost in training process util the coverence of loss on $PAMAP2$ without dropout. (d) Total time cost in training process util the coverence of loss on $PAMAP2$ with dropout.



(a) Number of data in storage on $HHAR$ (per thousand).

(b) Number of data in storage on $PAMAP2$ (per thousand).

Fig. 10. Storage cost with the increase of number of activities on two datasets: $HHAR$ and $PAMAP2$ during training. (a) Number of data in storage on $HHAR$ (per thousand). (b) Number of data in storage on $PAMAP2$ (per thousand).

stored, the model can still be updated to learn new activities incrementally. Compared with the batch training method, the storage cost can be greatly improved.

VI. Conclusion

Human activities recognition (HAR) based on multi-modality sensor data in incremental learning way was an important yet challenging task, which can save storage space in training process and easily scalable to new kind of activities comparing with traditional batch learning algorithms. In this paper, we proposed a multi-modality incremental learning model called $HarMI$ to address the problem. The $HarMI$ first adopted attention mechanism to align sensor data with very different frequencies to eliminate heterogeneity of different sensors. Then as catastrophic forgetting is common yet challenging in incremental learning, we overcome catastrophic forgetting from a multi-modality perspective based on elastic weight consolidation (EWC) framework by introducing the EWC regularization term and the correlation regularization term to preserve knowledge in previous activities. As demonstrated in the experiments, the proposed model $HarMI$ outperformed the state-of-the-art baselines on two public datasets. We also showed the superiority of $HarMI$ comparing with batching training methods in terms of time cost and storage cost, which envisions practical applications of $HarMI$ in resources-limited environments particularly.

The study of HAR in a multi-modality incremental way achieves several cheerful results. The proposed *HarMI* can mitigate performance degradation caused by catastrophic forgetting. However, there is still a performance gap between incremental learning models and batch training methods. In addition, it is easier to sample large number of unlabeled sensor data in real

world scenarios, which could be utilized to improve the model's performance based on semi-supervised learning. Furthermore, the sensor data produced by different individuals could be quite heterogeneous, which is non-independently identically distributed. How to eliminate the heterogeneity and utilize the similarity in incremental HAR is an interesting topic.

REFERENCES

[1] W. C. Abraham and A. Robins, "Memory retention ? the synaptic stability versus plasticity dilemma," *Trends Neurosci.*, vol. 28, no. 2, pp. 0–78, 2005.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[3] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[4] K. Chen *et al.*, "Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1–8.

[5] K. Chen, L. Yao, D. Zhang, B. Guo, and Z. Yu, "Multi-agent attentional activity recognition," In *Proc. 28th Int. Joint Conf. Artif. Intell. Organization*, Jul. 2019, pp. 1344–1350.

[6] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities," *ACM Comput. Survey (CSUR)*, vol. 54, no. 4, pp. 1–40, 2021.

[7] H. Dadkhahi and B. M. Marlin, "Learning tree-structured detection cascades for heterogeneous networks of embedded devices," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1773–1781.

[8] T. J. Draelos *et al.*, "Neurogenesis deep learning: Extending deep networks to accommodate new classes," in *2017 Int. Joint Conf. Neural Netw. (IJCNN)*, IEEE, pp. 526–533, 2017.

[9] C. Fernando *et al.*, "Pathnet: Evolution channels gradient descent in super neural networks," 2017, *arXiv:1701.08734*.

[10] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends Cogn. Sci.*, vol. 3, no. 4, pp. 128–135, 1999.

[11] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Inf. Fusion*, vol. 40, pp. 45–56, 2017.

[12] A. Gepperth and C. Karaoguz, "A. bio-inspired incremental learning architecture for applied perceptual problems," *Cogn. Comput.*, vol. 8, no. 5, pp. 924–934, 2016.

[13] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal, and J. Gama, "Machine learning for streaming data," *ACM SIGKDD Explorations Newslett.*, vol. 21, no. 2, pp. 6–22, 2019.

[14] B. Goodrich and I. Arel, "Unsupervised neuron selection for mitigating catastrophic forgetting in neural networks," in *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, 2014, pp. 997–1000.

[15] Y. Guan and T. Plötz, "Ensembles of deep lstm learners for activity recognition using wearables," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–28, 2017.

[16] V. Hernandez, T. Suzuki, and G. Venture, "Convolutional and recurrent neural network for human activity recognition: Application on american sign language," *PLoS One*, vol. 15, no. 2, 2020, Art. no. e0 228869.

[17] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," 2018, *arXiv:1802.02871*.

[18] H. Hotelling, "Relations between two sets of variates," *Breakthroughs in Statistics*. New York, NY: Springer, 1992, pp. 162–190.

[19] A. Jarraya, A. Bouzeghoub, A. Borgi, and K. Arour, "DCR: A new distributed model for human activity recognition in smart homes," *Exp. Syst. Appl.*, vol. 140, no. Feb., pp. 112849. 1–112849.19, 2020.

[20] K. Jia, J. Lin, M. Tan, and D. Tao, "Deep multi-view learning using neuron-wise correlation-maximizing regularizers," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5121–5134, Oct. 2019.

[21] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3390–3398.

[22] J. Kirkpatrick *et al.* "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.

[23] A. K. S. Kushwaha, S. Srivastava, and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," *Multimedia Syst.*, vol. 23, no. 4, pp. 451–467, 2017.

[24] O. D. Lara and M. A. Labrador, "A mobile platform for real-time human activity recognition," in *Proc. Consum. Commun. Netw. Conf.*, 2012, pp. 667–671.

[25] M. Lazarescu, S. Venkatesh, and G. West, "Incremental learning with forgetting," *Proc. ICML Workshop Mach. Learn. Comput. Vis.*, 1999, pp. [unknown]–[unknown].

[26] S.-W. Lee, C.-Y. Lee, D.-H. Kwak, J. Kim, J. Kim, and B.-T. Zhang, "Dual-memory deep learning architectures for lifelong learning of everyday human behaviors," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 1669–1675.

[27] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[28] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6467–6476.

[29] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition," in *Proc. 28th Int. Joint Conf. Artif. Intell*, 2019, pp. 3109–3115.

[30] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2001–2010.

[31] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.*, 2012, pp. 108–109.

[32] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 4255–4262.

[33] A. Stisen *et al.*, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in *Proc. 13th ACM Conf. Embedded Networked Sensor Syst.*, 2015, pp. 127–140.

[34] S. Sun and G. Chao, "Multi-view maximum entropy discrimination," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1706–1712.

[35] Y. Xu, E. Wang, Y. Yang, and Y. Chang, "A unified collaborative representation learning for neural-network based recommendersystems," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2021.

[36] Y. Xu, Y. Yang, E. Wang, F. Zhuang, and H. Xiong, "Detect professional malicious user with metric learning in recommender systems, *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2020.

[37] P. Yang and W. Gao, "Information-theoretic multi-view domain adaptation: A theoretical and empirical study," *J. Artif. Intell. Res.*, vol. 49, pp. 501–525, 2014.

[38] Y. Yang, D.-W. Zhou, D.-C. Zhan, H. Xiong, and Y. Jiang, "Adaptive deep models for incremental learning: Considering capacity scalability and sustainability," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 74–82.

[39] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 21–29.

[40] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 351–360.

[41] H. Ying *et al.*, "Sequential recommender system based on hierarchical attention networks," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3926–3932.

[42] M. Zeng *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput., Appl. Serv.*, 2014, pp. 197–205.

[43] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. 34th Int. Conf. Mach. Learn.*-Volume vol. 70, 2017, pp. 3987–3995.

[44] X. Zhang, F. Zhuang, W. Li, H. Ying, H. Xiong, and S. Lu, "Inferring mood instability via smartphone sensing: A multi-view learning approach," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1401–1409.

[45] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, 2017.

[46] Q. Zhou *et al.*, "Modeling heterogeneous relations across multiple modes for potential crowd flow prediction," vol. 35, no. 5, pp. 4723–4731, 2021.