A Semisupervised Recurrent Convolutional Attention Model for Human Activity Recognition

Kaixuan Chen, *Student Member, IEEE*, Lina Yao, *Member, IEEE*, Dalin Zhang,
Xianzhi Wang, *Member, IEEE*, Xiaojun Chang, and Feiping Nie

***Abstract*— Recent years have witnessed the success of deep learning methods in human activity recognition (HAR). The longstanding shortage of labeled activity data inherently calls for a plethora of semisupervised learning methods, and one of the most challenging and common issues with semisupervised learning is the imbalanced distribution of labeled data over classes. Although the problem has long existed in broad real-world HAR applications, it is rarely explored in the literature. In this paper, we propose a semisupervised deep model for imbalanced activity recognition from multimodal wearable sensory data. We aim to address not only the challenges of multimodal sensor data (e.g., interperson variability and interclass similarity) but also the limited labeled data and class-imbalance issues simultaneously. In particular, we propose a pattern-balanced semisupervised framework to extract and preserve diverse latent patterns of activities. Furthermore, we exploit the independence of multi-modalities of sensory data and attentively identify salient regions that are indicative of human activities from inputs by our recurrent convolutional attention networks. Our experimental results demonstrate that the proposed model achieves a competitive performance compared to a multitude of state-of-the-art methods, both semisupervised and supervised ones, with 10% labeled training data. The results also show the robustness of our method over imbalanced, small training data sets.**

***Index Terms*— Attention, class imbalance, human activity recognition (HAR), semisupervised learning.**

I. Introduction

HUMAN Activity Recognition (HAR) is a fundamental technique popular in healthcare and surveillance domains [1]. In particular, wearable physical sensor signal processing-based HAR has been widely applied to ubiquitous applications and profoundly revolutionized our daily lives, thanks to its high resistance to environmental variation without significantly violating individual privacy.

Although remarkable efforts have been contributed to different aspects of HAR, three challenges remain for the research community. The first is insufficient labeled observations [2]. Most existing works follow a supervised learning approach [3], [4], thus requiring a significant amount of training data to recognize meaningful activities. However, ground truth annotation is usually both costly and error-prone. Semisupervised methods, in contrast, additionally leverage unlabeled data to train the model and therefore are considered more promising in many scenarios. Although researchers have already investigated several semisupervised techniques [5], [6], they neglect the benefit of combining multimodality sensor data and overlook the inner patterns of each activity. Considering its superiority in dealing with multimodalities, we resort to a cotraining method [7]. Previous studies [8] suggest that cotraining algorithms can work well when the multiple views (or multimodalities) are not strongly correlated [9], with each view containing sufficient information to learn a weakly useful classifier and other views redundant for this view [7]. The above property makes cotraining appropriate for the multimodal activity recognition problem.

The second challenge concerns the expense and convenience of labeling activity data, where the class imbalance is often a concurrent issue. Especially in HAR tasks, some activity data (e.g., those related to falls of elder people) are difficult to obtain and label. In fact, semisupervised learning on imbalanced classification is even more challenging. While most classifiers tend to predict majority class samples with high accuracy and treat the minority classes as outliers [10], the situation becomes more severe when only a small amount of data is available. Previous works directly apply undersampling or oversampling [11], but they are unsuitable for our case as they both change the distribution of the training data. Since, in our case, the same individual may perform the same activity in different ways because of stress, fatigue, emotion, and other environmental factors, it is reasonable to assume that samples in each class can form several latent patterns. Therefore, we select the training samples in each training round in line with the extracted latent patterns to sustain the diversity of activity patterns. Such a pattern-preserving

Manuscript received September 20, 2018; revised March 23, 2019 and May 30, 2019; accepted July 3, 2019. *(Corresponding author: Kaixuan Chen.)*

K. Chen, L. Yao, and D. Zhang are with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: kaixuan.chen@student.unsw.edu.au; lina.yao@unsw.edu.au; dalin.zhang@student.unsw.edu.au).

X. Wang is with the School of Computer Science, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: xianzhi.wang@uts.edu.au).

X. Chang is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: cxj273@gmail.com).

F. Nie is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2927224

framework maintains the distribution of training data and improves the labeling performance during training.

The third challenge contains two parts and is longstanding for HAR: *interperson variability* and *interclass similarity* [2]. The interperson variability means the same activity can be performed differently by different people and interclass similarity results from the similarity in the behavior patterns of different activities such as walking and running. Since deep learning based methods have the strength of modeling the high-level representations of data and have achieved outstanding performance in the field of HAR, we aim to explore more potential of deep learning models in the field of HAR. We train "attentive" deep models to extract the salient information indicative of the true activity to obtain useful information from limited training data and address the influences of the inter-person variability and interclass similarity.

In brief, we deploy a pattern-balanced recurrent convolutional attention model to address the above concerns. Our approach achieves high accuracy on a small size, imbalanced data. To the best of our knowledge, this paper is the first that uses semisupervised learning in imbalanced activity recognition. The key contributions of this research are as follows.

1) We propose a novel method that employs semisupervised learning for imbalanced HAR, which is a significant challenge yet rarely explored in the literature.
2) Considering the influence of class imbalance on limited training data, we propose a pattern-balanced cotraining for extracting and preserving the latent activity patterns from imbalanced data sets. The patterns maintain the distribution of training data and improve the robustness of cotraining on imbalanced data.
3) To better utilize the limited labeled data and get higher labeling accuracy during training, we employ Recurrent Attention Models (RAMs) [12] and let them collaborate to exploit unlabeled samples.
4) We compare our model with the state-of-the-art methods on three public benchmarked data sets and a new data set collected in the real world. The experimental results show that our approach achieves competitive performance compared to the state-of-the-art semisupervised and even supervised methods with 10% labeled training data.

II. Related Work

A. semiSupervised Learning for Imbalanced HAR

Although deep learning methods achieve high recognition performance in HAR [4], they require a substantial amount of labeled activity data for training. semisupervised learning allows leveraging both labeled and unlabeled data to train a HAR system [1]. Some works use self-learning-based approaches [6] and some utilize graph-based approached [5], [13]. Nevertheless, these approaches all rely on *ad hoc* hand-crafted features, which makes it hard and expensive to build a recognition system. [14], [15] resort to deep generative models such as Restricted Boltzmann Machines (RBMs) and autoencoders to train the model with a significant amount of unlabeled data and get a well-trained feature extractor. The feature extracted can be further recognized with classifiers trained with labeled data. Methods like these, however, suffer from three defects. First, none of these methods takes advantage of the relations between multimodalities of activity data which is of great significance since modalities carry information from different perspectives and complement each other. Second, they fail to explore the potential activity category information of those unlabeled data [4]. The reason why discriminative models cannot directly participate in semisupervised learning is that the small size of labeled data is not enough for training. In contrast, our method utilizes the disagreement between modalities and introduces attention mechanisms, so deep learners can learn to exploit potential class information of unlabeled samples with small labeled sets.

However, none of these semisupervised methods consider a more common case where the class distribution is imbalanced. Although some approaches [10], [16] have been proposed to solve the imbalance issue, their performance deteriorates due to the small data size in semisupervised learning. The works in [11], [17] are devoted to semisupervised learning for imbalanced learning, yet they aim at solving a binary classification problem, which is more straightforward than recognizing multiple activities in HAR. So far, semisupervised learning for imbalanced HAR problems still has not been carefully studied in the literature.

B. Attention Mechanisms

Attention originates from biology and psychology that implies focusing the power of noticing or thinking on something special to achieve better cognitive processes. Tracing back the history of selecting effective regions using attention mechanisms or similar theories, attention-based recurrent neural network (RNN) model has achieved success in both speech recognition [18] and computer vision [12], [19]. Bahdanau *et al.* [18] build a vocabulary continuous speech recognition system using an attention-based RNN as it requires fewer training stages, fewer auxiliary data, and less domain expertise. Some works [12], [19] formulate the selection process into a sequential decision task. Our previous work [20], [21] adopts the attention mechanism for HAR. We fuse attention with convolutional neural network (CNN) and RNN to automatically extract the most salient modality-specific features and further convert the information to higher level representation. In this work, our approach allows the attention mechanism to fully leverage its strengths to strive for a balance from less labeled data.

III. Methodology

A. Overview

In this section, we propose an integrated system for semisupervised and imbalanced HAR. First, we propose a pattern-balanced framework that preserves and balances diverse intent patterns of activities. The proposed framework improves the performance of conventional cotraining under imbalanced labeled data (Section III-B). Second, considering the limited labeled data, we aim at maximizing the utilization of salient

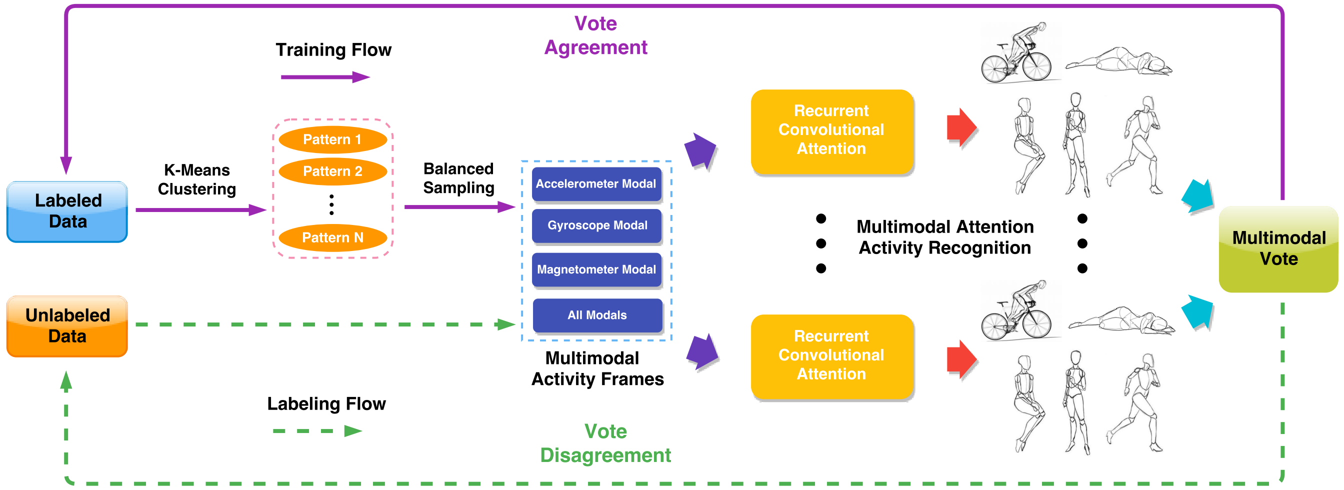


Fig. 1. Workflow of the proposed pattern-balanced co-training framework. The framework contains two flows. 1) Training flow (solid lines): labeled data are categorized into N patterns via k-means clustering and data of patterns are sampled evenly to train multimodal classifiers. 2) Labeling flow (dashed lines): predict the unlabeled data with trained models. If most of the classifiers reach an agreement on predicting a sample, this sample is labeled; otherwise, keep it unlabeled.

features and ignoring the irrelevant signals. We introduce attention mechanism and deploy recurrent convolutional attention models to get better labeling accuracy on limited labeled data and deal with the interperson variability and interclass similarity of HAR (Section III-C).

Since the modalities in our case satisfy the sufficiency, redundancy, and weak relations that cotraining requires, we develop an effective semisupervised framework based on cotraining method [7] to handle limited labeled data.

The basic framework of cotraining is as follows.

1) The training data contains two parts, namely, labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$.
2) Three classifiers are trained on $\mathcal{L}$ of acceleration, angular velocity, and magnetism, respectively.
3) Each model is applied to $\mathcal{U}$ to make a prediction and vote to label the most confident samples. These selected samples are removed from $\mathcal{U}$ and added to $\mathcal{L}$ to improve the classifiers in the following training rounds.
4) Repeat steps 2) and 3) until no more samples can be voted or $\mathcal{U}$ turns empty.
5) Train a classifier with the final $\mathcal{L}$ and all modalities. Owing to the prevalence and excellent performance of deep learning methods in HAR, we deploy deep learning classifiers and embed them into the cotraining framework.

Although the framework above seems to be valid, we observe that the labeling accuracy decreases with training rounds since the problem of class imbalance is severe when the labeled data size is small. The practical difficulty of obtaining and labeling some specific activity data (e.g., falls of elder people) make the problem even more challenging. Therefore, we propose to tackle the class imbalance by pattern-balanced training in Section III-B.

B. Class Imbalance Mitigation

Our proposed pattern-balanced training is robust to class-imbalanced labeled data. Fig. 1 shows the overall workflow of our framework. The workflow contains a training flow and a labeling flow.

1) Training Flow: Its goal is to train weak classifiers with labeled data so that they can vote to label samples from unlabeled data with high accuracy even in class-imbalanced situations. For clarity, the labeled set $\mathcal{L}$ contains L labeled samples. Each sample $(\mathbf{x}, y)$ consists of a vector that represents the collected sensory data $\mathbf{x}$ and the activity label y. Since most inertial measurement units (IMUs) used in HAR community contains three inertial sensors, namely, an accelerometer, a gyroscope, and a magnetometer, we suppose that

$$\mathbf{x} = (\mathbf{Acc}, \mathbf{Gyro}, \mathbf{Magn}) \tag{1}$$

and

$$y \in [1 \ldots C] \tag{2}$$

where C denotes the number of activity classes. To focus on the class imbalance problem, $\mathcal{L}$ is separated to C classes according to their labels so that

$$\mathcal{L} = \bigcup_{i=0}^{C} \mathcal{C}_i \tag{3}$$

where

$$\mathcal{C}_i = \{(\mathbf{x}, y) \mid y = i\} = \{(\mathbf{x}_0, i), \ldots (\mathbf{x}_j, i) \ldots (\mathbf{x}_{\mathrm{Card}(\mathcal{C}_i)}, i)\} \tag{4}$$

and $Card(\mathcal{C}_i)$ represents the cardinality of $\mathcal{C}_i$.

A basic solution to class imbalance is to oversample small-class data or undersample large-class data [22], but it may change the distribution of training data and lead to “covariate shift.” On the other hand, considering the intraclass variation of HAR, we aim at preserving the diversity of patterns within each class to avoid the distribution shift. As the latent patterns rely on expertise, we adopt k-means clustering to each class $\mathcal{C}_i$ to extract activity patterns by minimizing the measurement:

$$D = \sum_{m=1}^{\mathrm{Card}(\mathcal{C}_i)} \sum_{k=1}^{K_i} \boldsymbol{\mu}_{\mathrm{km}} \|\mathbf{x}_m - \mathbf{z}_k\| \tag{5}$$

where K_i denotes the number of clusters of $\mathcal{C}_i$ and it is adaptively decided by its covariance [23]. $\boldsymbol{\mu}_{km} = 1$ if $\mathbf{x}_m$ belongs to the cluster $\mathcal{Z}_k$ with center $\mathbf{z}_k$; otherwise, $\boldsymbol{\mu}_{km} = 0$. Therefore,

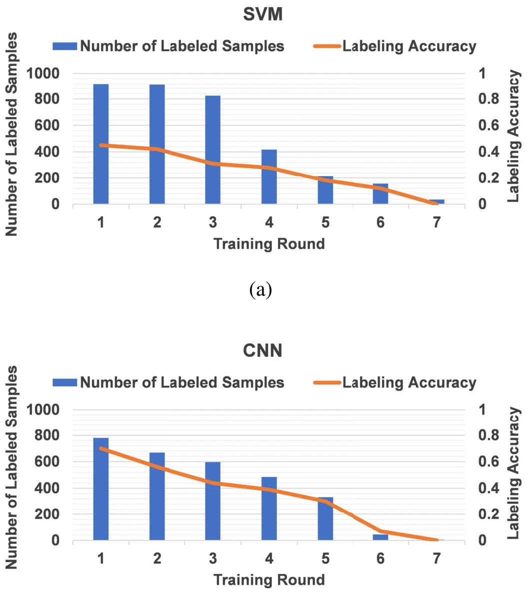
$$\mathcal{C}_i = \bigcup_{n=0}^{K_i} \mathcal{Z}_n. \tag{6}$$

After extracting the activity patterns $\mathcal{Z}$, we apply oversampling and undersampling to classes to maintain the distribution. In particular, we randomly select the same number of samples from patterns $\mathcal{Z}$ of classes $\mathcal{C}$ to make sure that samples from all patterns evenly participate in the next training round. We conduct clustering and sampling in the training flow rather than the labeling flow to avoid repeatedly labeling samples. Then, four classifiers are trained separately for acceleration, angular velocity, magnetism, and the combination of all modalities. All modalities are treated as the fourth modality to guarantee the labeling accuracy.

2) Labeling Flow: In the labeling flow, the trained classifiers are applied to unlabeled data to make a prediction. Four classifiers vote to select confident samples and label them. These selected samples are removed from the unlabeled set $\mathcal{U}$ and added to $\mathcal{L}$. Next, the training flow and the labeling flow are repeated until no confident samples can be labeled or $\mathcal{U}$ is empty. The last step is to train a classifier on the final labeled set with all modalities. In our framework, we only fine-tune the fourth classifier (which is repeatedly trained in the training flow) with the newly labeled samples. The target of the labeling flow is to leverage the sufficiency and redundancy of multi-modalities so that four classifiers can learn from each other.

C. Limited Data Exploitation

Even with pattern-balanced training, another issue hinders the HAR concerning imbalanced, small labeled sets. After the sampling, the already limited labeled set becomes even smaller. As a result, it is hard to train satisfactory models with so limited labeled data in our case, especially when HAR data naturally suffers from interperson variability and interclass similarity. Fig. 2 shows the labeling accuracy and the numbers of samples labeled in each training round with 2000 balanced labeled data and 18 000 unlabeled data. First, CNN enjoys the merit of deep learning, so the labeling accuracy of CNN is higher than that of support-vector machine (SVM) in the first round. However, as the models only have 2000 original labeled samples, the labeling accuracy is low (0.45 for SVM and 0.72 for CNN). As SVM labeled about 900 samples with 0.45 accuracy, which means they introduce 550 falsely labeled samples to $\mathcal{L}$ approximately. The falsely labeled samples create a vicious circle and further decreases the labeling accuracy in the following training rounds until the accuracy is 0. With respect to CNN, as CNN has a relatively higher accuracy in the first round, it has a more stable decrease in both the labeled data number and the labeling accuracy, but still cannot avoid the continuously decreasing accuracy. Therefore, it is necessary to train classifiers that can fully exploit salient features from limited data and achieve high labeling accuracy. In this section, we use RAMs [12] which extract the informative features. They can learn modality-specific information and distinguish disagreement among the modalities so that these models can learn from each other and do not incorrectly vote samples to the same labels.



(a)

(b)

Fig. 2. Labeling accuracy and numbers of labeled samples versus training rounds on 2000 original labeled data and 18 000 unlabeled data. Orange lines: labeling accuracy. Blue bars: numbers of samples labeled in each training round. (a) SVM. (b) CNN.

1) Motivations: Intuitively, the motion of different body parts has various contributions to different activities [24]. For example, jumping involves legs while running is related to both arms and legs; another example is that recognizing patterns of walking depends more on the acceleration of legs while distinguishing sitting from lying would rely more on the orientation of sensor placement. With these characteristics of motion data considered, the natural idea is to focus on the most highly contributing part of several modality data. Inspired by the procedures of human brains processing visual information, we introduce the attention mechanism into HAR systems. RAM is particularly efficient in the scenario when the number of labeled data is limited as it maximizes the effect of useful information and alleviates the influence of interperson variability and interclass similarity. Hence, it can work effectively even when the labeled data size is small.

Fig. 3 shows the basic structure, where the whole process sits on a core long short-term memory (LSTM). At each time step t, the model only focuses on a small patch which is called a glimpse. To extract the most salient patch, we train the model

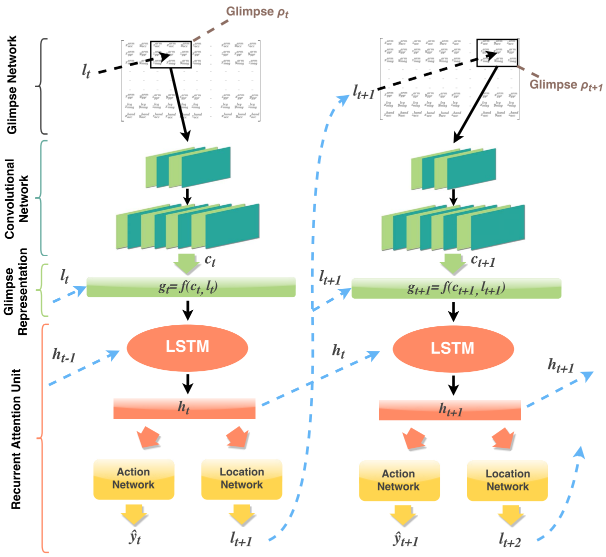


Fig. 3. RAM. The input data are represented as matrices. At each time step t, a small glimpse patch is extracted with the glimpse layer and processed by a convolutional network. The processed glimpse information is then encoded with location information in the following step. A recurrent network with two sub-networks predicts the activity at the current time step and decide the glimpse location for the next time step.

using reinforcement learning [12]. RAM consists of four key components: 1) a glimpse layer; 2) a convolutional network; 3) a glimpse representation layer; and 4) a core recurrent attention unit. We explain the details in the following.

2) Glimpse Layer: The first part of the proposed model is a glimpse layer. The glimpse layer not only avoids the system processing the whole data in their entirety but also maximally eliminates the information loss raised by traditional dimensionality reduction and feature selection [12], [19]. As sensory data do not have fixed ordering arrangement, we preprocess the multimodal sensor data by transforming them from sequences to matrices with the arrangement algorithm proposed in our previous work [20]. This arrangement extracts the full correlations between feature pairs so that the glimpses selected may contain relations between both adjacent and nonadjacent features. Inspired by the human visual system, in RAM, each input matrix I will be "understood" within T glimpses. Simulating the process of how the human eye works, RAM extracts a glimpse region denoted by ρ_t from the input matrix I at the location l_t at each time step t.

3) Convolutional Network: Human visual system converts retina images into brain signals via the optic nerves. Likewise, we convert the glimpse directly extracted from the input matrix into higher-level information. RAM uses a convolutional network to encode ρ_t to be c_t, parameterized by θ_c, which generates a high-level representation that characterizes the local salience of the low-level sensor data:

$$c_t = \mathrm{Conv}(\rho_t; \theta_c). \tag{7}$$

4) Glimpse Representation Layer: The glimpse needs to be further processed by a glimpse representation layer [12], [19]. c_t and the location l_t are linearly transformed independently with two linear layers parameterized by θ_g^c and θ_g^l, respectively. Next, the summation of these two parts is further transformed with another linear layer parameterized by θ_g and a rectified linear unit. The whole process is summarized as follows:

$$g_t = f_g\left(c_t, l_t; \theta_g^c, \theta_g^l, \theta_g\right) = \mathrm{relu}(L(L(c_t) + L(l_t))) \tag{8}$$

where $L(\bullet)$ denotes a linear transformation. Therefore, the glimpse representation g_t finally contains information from both "what" (c_t) and "where" (l_t).

5) Recurrent Attention Unit: We use a RNN as the core to process data step by step within several glimpses. As shown in Fig. 3, the basic structure of the recurrent attention unit is an LSTM. At each time step t, the LSTM receives the glimpse g_t and the previous hidden state h_{t-1} as the inputs parameterized by θ_h. Meanwhile, it outputs the current hidden state h_t according to the equation:

$$h_t = f_h(h_{t-1}, g_t; \theta_h). \tag{9}$$

The recurrent attention unit also contains two subnetworks: the *location* network and the *action* network. These two subnetworks receive the hidden state h_t as the input to decide the next glimpse location l_{t+1} and the current action a_t. The location network outputs the location at time $t+1$ stochastically according to the location policy defined by a Gaussian distribution stochastic process, parameterized by the location network $f_l(h_t; \theta_l)$

$$l_{t+1} \sim P(\cdot \mid f_l(h_t; \theta_l)). \tag{10}$$

Similarly, the action network outputs the corresponding action at time t and predicts the activity label $\hat{y}_t$ given the hidden state h_t. The action $\hat{y}_T$ at the last time step T indicates the final prediction of the activity. $\hat{y}_t$ obeys the distribution parameterized by $f(h_t; \theta_a)$. Due to its prediction function, the network uses a softmax formulation

$$a_t = \hat{y}_t = f_a(h_t; \theta_a) = \mathrm{softmax}(L(h_t)). \tag{11}$$

6) Training and Optimization: Our proposed model involves the parameters of the convolutional network, the glimpse representation layer and the two sub-networks, $\Theta = \theta_c, \theta_g^c, \theta_g^l, \theta_g, \theta_h, \theta_a, \theta_l$. Since the action network relies on classification methods, θ_a can be trained by optimizing the cross-entropy loss and the backpropagation. However, we expect the location network to be able to select a sequence of salient regions from input matrices adaptively. In view that the location network is stochastic and nondifferentiable, the salient region selection problem can also be seen as a control problem, and it can be trained by reinforcement methods to learn the optimal policies. Based on the above discussion, we deploy a Partially Observable Markov Decision Process (POMDP) to solve the training and optimization problems [25]. In particular, we call the sequence of the input, location, and action pairs, $s_{1:t} = \mathbf{x}, l_1, \hat{y}_1; \ldots \mathbf{x}, l_t, \hat{y}_t$, an attention sequence and use this sequence to represent the order of the regions that the model focused on. In our case, the location network is formulated as a random stochastic process (the Gaussian distribution) parameterized by Θ. Each time after the location selection, the prediction $\hat{y}$ is evaluated to back feed a reward r for conducting the backpropagation training process. $r_t = 1$ if $\hat{y}_t = y_t$ and 0 otherwise. The process is also defined as policy gradient. Our goal is to maximize the simulated rewards R using gradient. Given a sample x with reward $f(x)$ and probability $p(x)$, the gradient can be calculated as follows:

$$\begin{aligned}\nabla_\theta E_x[f(x)] &= \nabla_\theta \sum_x p(x) f(x) \\ &= \sum_x \nabla_\theta p(x) f(x) \\ &= \sum_x p(x) \frac{\nabla_\theta p(x)}{p(x)} f(x) \\ &= \sum_x p(x)(\nabla_\theta \log p(x)) f(x) \\ &= E_x[f(x)(\nabla_\theta \log p(x))]. \qquad (12)\end{aligned}$$

In our case, given the reward R and the attention sequence $s_{1:T}$, the reward function to be maximized is as follows:

$$J(\Theta) = \mathbb{E}_{p(s_{1:T};\Theta)}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}_{p(s_{1:T};\Theta)}[R]. \qquad (13)$$

By considering the training problem as a POMDP, a sample approximation to the gradient is calculated as follows according to the REINFORCE rule [26]:

$$\nabla_\Theta J = \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:t};\Theta)}[R \nabla_\Theta \log\pi(y|s_{1:t};\Theta)]. \qquad (14)$$

We use Monte Carlo sampling which utilizes randomness to yield results that might be deterministic theoretically. Supposing M is the number of Monte Carlo sampling copies, we duplicate the same input for M times and average them as the prediction results to overcome the randomness in the network, where the M duplication generates M subtly different results owing to the stochasticity, we have

$$\nabla_\Theta J \approx \frac{1}{M} \sum_{i=1}^{M} R^{(i)} \sum_{t=1}^{T} \nabla_\Theta \log\pi\left(y^{(i)}|s^i_{1:t};\Theta\right) \qquad (15)$$

where i denotes the ith training sample, $y^{(i)}$ is the correct label for the ith sample, $\nabla_\Theta log\pi(y^{(i)}|s^i_{1:t};\Theta)$ is the gradient of LSTM calculated by backpropagation, and M denotes the number of Monte Carlo sampling copies used for overcoming the randomness of the networks.

Therefore, although the best attention sequences are unknown, RAM can learn the optimal policy in the light of the reward. The experiments show that RAM outperforms the state-of-the-art in the initial phase of cotraining.

IV. Experiments

In this section, we evaluate the proposed method on four data sets. Three of them, MHEALTH [30], PAMAP2 [31], and UCI HAR [32], are public benchmarked data sets on activity recognition. They are the latest available multimodal wearable sensor-based data sets with complete annotation. The other one, Multimodal Activity Recognition with Sensing (MARS), is a real-world data set which we collected to reexamine the practicability of the proposed method. This data set is collected while eight participants (six males, two females) are performing five basic activities (sitting, standing, walking, ascending stairs, and descending stairs).

We first compare our method with different state-of-the-art works and baselines under both supervised and semisupervised schemes. Then, we explore the robustness to class imbalance by comparing the proposed method with state-of-the-art and baselines in five class imbalance situations. Third, we perform a detailed ablation study to examine the contributions of the proposed components to the prediction performance. Finally, we visualized the selected features by the attention model. Due to the page constraint, some details and experimental results are presented in the supplementary materials. The materials contain model implementation, confusion matrices, time latency comparison, empirical studies on annotation scarcity and class imbalance, analysis of training evolution, and hyperparameter study.

A. Robustness to Annotation Scarcity

To verify our semisupervised approach's robustness to data scarcity, we extensively compare our model with a set of state-of-the-art and baseline methods trained with different numbers of labeled samples. The compared methods include both supervised approaches and semisupervised approaches. We compare our approach with supervised approaches not only to exhibit the robustness to annotation scarcity but also to show that our classifiers are stronger than these supervised methods.

Table I presents the comparison between the proposed approach and the state-of-the-art methods as well as baselines. The data sets used in these experiments are class balanced. The notation "sup" indicates supervised methods while "semisup" indicates semisupervised methods. For fairness, we only deploy the supervised methods with sufficient labeled data. Similar to [3], the experiments conducted on the PAMAP2 and MHEALTH data sets perform background activity recognition tasks [31]. The activities belong to six classes: lying, sitting/standing, walking, running, cycling, and other activities. All the models are implemented on the above four data sets with parameters either indicated in the literature or via careful parameter tuning. We can observe that from 1000 labeled data to 2000 labeled data, there is a relatively large gap. However, the performance of the model trained with 2000 labeled samples is only slightly worse than that of the model trained with 20 000 labeled samples. Moreover, with only 2000 labeled samples and 18 000 unlabeled samples, our approach achieves competitive or better results than the supervised methods. Even with only 1000 samples, the results are acceptable. Compared to the other semisupervised methods, the proposed method shows significant improvement (at least 9%), which demonstrates the effectiveness of the pattern-balanced cotraining framework. Also, when we train RAM in a supervised manner (i.e., with 20 000 labeled samples), the results are better than the state-of-the-art supervised methods, indicating that RAM outperforms the other methods.

B. Robustness to Class Imbalance

We conduct experiments in five class imbalance situations to explore the robustness to class imbalance. As the

TABLE I

CLASSIFICATION ACCURACY OF PROPOSED APPROACH AND SIX STATE-OF-THE-ART METHODS AND BASELINE METHODS ON DIFFERENT SIZES OF LABELED SETS. THE NUMBERS OF THE LABELED DATA USED BY THE SEMISUPERVISED MODELS ARE DENOTED IN EACH COLUMN (X) AND THE NUMBERS OF THE UNLABELED SAMPLES ARE 20 000-X. * INDICATES OUR APPROACH WITH 2000 LABELED DATA OUTPERFORMS OR IS COMPETITIVE WITH THE COMPARED METHODS WITH FULL LABELED DATA

Dataset	Method	Training Scheme	Labeled Samples							
			1000	2000	5000	7000	10000	13000	17000	20000
MHEALTH	Multichannel CNN [4]	Sup	0.6226	0.6285	0.6936	0.7108	0.7573	0.8304	0.8587	0.8719
	Attention [20]	Sup	0.8152	0.8573	0.8970	0.9002	0.9178	0.9149	0.9242	0.9392
	Modality-Specific [27]	Sup	0.6358	0.6655	0.7616	0.7697	0.8336	0.8486	0.8620	0.8967
	Co-Training+CNN	Semi-Sup	0.6331	0.6495	0.6538	0.7029	0.7932	0.8110	0.8503	0.8604
	Diversity Preserving [28]	Semi-Sup	0.7023	0.7572	0.8271	0.8535	0.8721	0.8789	0.8892	0.8904
	Tri-Net [29]	Semi-Sup	0.6938	0.6963	0.7288	0.8076	0.8582	0.8603	0.8661	0.8754
	Our Approach	Sup	0.8059	0.8327	0.8714	0.8924	0.8935	0.9194	0.9326	0.9405
	Our Approach	**Semi-Sup**	**0.8895**	**0.9194**	**0.9208**	**0.9264**	**0.9385**	**0.9425**	**0.9411**	**0.9405**
PAMAP2	Multichannel CNN [4]	Sup	0.5405	0.6255	0.6435	0.6483	0.7167	0.7400	0.7976	0.8116
	Attention [20]	Sup	0.6443	0.7484	0.7756	0.7785	0.7869	0.8045	0.8187	0.8239
	Modality-Specific [27]	Sup	0.5578	0.5641	0.6169	0.6949	0.7728	0.7949	0.8040	0.8208
	Co-Training+CNN	Semi-Sup	0.5998	0.6042	0.6321	0.6612	0.6857	0.7328	0.7768	0.7922
	Diversity Preserving [28]	Semi-Sup	0.6412	0.6471	0.6783	0.7394	0.7729	0.7861	0.7913	0.8023
	Tri-Net [29]	Semi-Sup	0.6329	0.6429	0.6541	0.6954	0.7252	0.7624	0.7955	0.8088
	Our Approach	Sup	0.6289	0.7305	0.7654	0.7749	0.7935	0.8069	0.8228	0.8342
	Our Approach	**Semi-Sup**	**0.7338**	**0.8125**	**0.8137**	**0.8135**	**0.8204**	**0.8329**	**0.8318**	**0.8342**
UCI HAR	Multichannel CNN [4]	Sup	0.5355	0.5531	0.5584	0.5696	0.6724	0.7368	0.7469	0.7586
	Attention [20]	Sup	0.6683	0.6969	0.6996	0.7104	0.7297	0.7733	0.8073	0.8129
	Modality-Specific [27]	Sup	0.5465	0.5502	0.5879	0.6269	0.6785	0.7360	0.7582	0.7753
	Co-Training+CNN	Semi-Sup	0.5212	0.5739	0.6215	0.7106	0.7059	0.7248	0.7201	0.7336
	Diversity Preserving [28]	Semi-Sup	0.6113	0.6502	0.7054	0.7008	0.7129	0.7310	0.7316	0.7408
	Tri-Net [29]	Semi-Sup	0.6156	0.6446	0.6607	0.6675	0.6852	0.7013	0.7207	0.7365
	Our Approach	Sup	0.6427	0.6863	0.6981	0.7124	0.7249	0.7554	0.7804	0.8132
	Our Approach	**Semi-Sup**	**0.7281**	**0.7762**	**0.7818**	**0.7851**	**0.8073**	**0.8143**	**0.8084**	**0.8132**
MARS	Multichannel CNN [4]	Sup	0.6628	0.6699	0.6868	0.7029	0.7351	0.7559	0.7921	0.8138
	Attention [20]	Sup	0.7223	0.7832	0.8018	0.8357	0.8434	0.8408	0.8497	0.8538
	Modality-Specific [27]	Sup	0.6751	0.6787	0.6867	0.6884	0.7009	0.7334	0.7680	0.8354
	Co-Training+CNN	Semi-Sup	0.6442	0.6538	0.6968	0.7162	0.7321	0.7259	0.7954	0.8125
	Diversity Preserving [28]	Semi-Sup	0.7158	0.7370	0.8157	0.8294	0.8238	0.8169	0.8191	0.8208
	Tri-Net [29]	Semi-Sup	0.6904	0.7084	0.7376	0.7594	0.7691	0.7857	0.8052	0.8193
	Our Approach	Sup	0.7058	0.7444	0.7689	0.7841	0.7902	0.8173	0.8318	0.8592
	Our Approach	**Semi-Sup**	**0.8041**	**0.8325**	**0.8429**	**0.8416**	**0.8457**	**0.8393**	**0.8364**	**0.8592**

TABLE II

FIVE CLASS IMBALANCE SITUATIONS OF EXPERIMENTS. $S1$ IS THE BASELINE SITUATION CONTAINING EVENLY DISTRIBUTED CLASSES. $S2$–$S5$ ARE FOUR SITUATIONS WHERE DATA OF THREE CLASSES ARE REDUCED TO 2000 WHILE DATA OF THE REST CLASSES ARE INCREASED TO 7000

	Classes	S1	S2	S3	S4	S5
Class Distribution	Sitting	4000	2000	7000	7000	2000
	Standing	4000	2000	2000	7000	7000
	Walking	4000	2000	2000	2000	7000
	Ascending Stairs	4000	7000	2000	2000	2000
	Descending Stairs	4000	7000	7000	2000	2000

background activity class "others" makes a considerable impact on HAR task, we filter the "other" class and perform five-class classification. We design five class-imbalance situations as shown in Table II. $S1$ is the basic situation containing evenly distributed classes. $S2$–$S5$ are four situations where data of three classes are reduced to 2000 while data of the rest classes are increased to 7000. Note that MHEALTH does not include ascending and descending stairs, so we replace them with cycling and climbing stairs. In addition to three state-of-the-art methods [16], [11], [10], we also compare our method with two baselines, Over-Sampling and Under-Sampling that perform sampling by randomly selecting or filtering out a certain number of samples. For fairness, we use RAMs as the classifiers for comparison.

Table III shows the performance of these methods on four data sets in five class situations. As F1 score is the most suitable measurement for class imbalanced problems, we use the F1 score in this table. After eliminating the impact of "others", the overall classification performance on MHEALTH and PAMAP2 is boosted. Since $S1$ enjoys even class distribution, we can observe that the overall performance in $S1$ is higher than that in other situations. Besides, both the sampling models and subspace generation use RAMs and achieve the same performance as ours in the balanced situation, because the classifiers are regularly trained without any strategy. Another observation is that the results in $S2$ are relatively low because the two activities, ascending and descending stairs, are hard to be distinguished even though there are plenty of data.

Regarding the methods, our approach outperforms the others in all situations, and the difference between the performance in class balanced situation and class imbalanced situations is not apparent. Among the other compared methods, sampling models and subspace generation have the most similar performance to our model owing to the same classifiers that they use, and subspace generation is more robust on imbalanced data. It is hard to distinguish which sampling strategy is better: undersampling simply throws out the information while oversampling may repeat some samples many times and lead

TABLE III

F1 Scores of All Compared Methods on Five Different Class Situations. The Details About $S1$–$S5$ Are Listed in Table II

	MHEALTH						PAMAP2					
	Over-Sampling	Under-Sampling	One-Class [16]	Subspace Generation [11]	Ensemble LSTM [10]	**Our Approach**	Over-Sampling	Under-Sampling	One-Class [16]	Subspace Generation [11]	Ensemble LSTM [10]	**Our Approach**
S1	0.9921	0.9921	0.7975	0.9921	0.9338	**0.9921**	0.9287	0.9287	0.7320	0.9287	0.8704	**0.9287**
S2	0.9425	0.9580	0.7630	0.9627	0.9052	**0.9832**	0.8818	0.8679	0.7064	0.9024	0.8350	**0.9207**
S3	0.9672	0.9659	0.7712	0.9715	0.9078	**0.9854**	0.8858	0.8746	0.7152	0.9134	0.8435	**0.9143**
S4	0.9608	0.9605	0.7756	0.9874	0.9184	**0.9916**	0.8903	0.8867	0.7281	0.9157	0.8361	**0.9141**
S5	0.9590	0.9735	0.7673	0.9824	0.9139	**0.9834**	0.8910	0.8939	0.7176	0.9093	0.8526	**0.9223**
	UCI HAR						MARS					
	Over-Sampling	Under-Sampling	One-Class [16]	Subspace Generation [11]	Ensemble LSTM	**Our Approach**	Over-Sampling	Under-Sampling	One-Class [16]	Subspace Generation [11]	Ensemble LSTM	**Our Approach**
S1	0.7212	0.7212	0.6255	0.7212	0.6804	**0.7212**	0.8437	0.8437	0.7383	0.8437	0.8273	**0.8437**
S2	0.6797	0.6615	0.5619	0.7026	0.6641	**0.7013**	0.8097	0.8114	0.6983	0.8158	0.8002	**0.8312**
S3	0.6820	0.6959	0.5844	0.7060	0.6613	**0.7191**	0.8265	0.8201	0.6943	0.8205	0.7945	**0.8349**
S4	0.6947	0.6967	0.6086	0.7095	0.6820	**0.7099**	0.8090	0.8219	0.6986	0.8377	0.8133	**0.8314**
S5	0.6836	0.6732	0.5927	0.7048	0.6644	**0.7159**	0.8038	0.8235	0.6925	0.8283	0.8079	**0.8564**

TABLE IV

Ablation Study. The Table Presents the Accuracy, Training Time (s), and Test Time (ms) of the Models on 2000 Balanced Training Data and the F1 Scores on 5000 Imbalanced Training Data. The Imbalanced Ratio Is as S2. The Training Time Is Shown in the Parenthesis

Ablation	Datasets	Balanced	Imbalanced	Test Time (ms)	Datasets	Balanced	Imbalanced	Test Time (ms)
RAM	MHEALTH	0.8327 (121.84)	0.8329 (182.98)	24.0	UCI HAR	0.6863 (94.31)	0.6084 (124.78)	19.7
Co-Training		0.6495 (127.42)	0.7208 (142.60)	8.9		0.5739 (116.94)	0.552 (121.28)	7.6
Pattern-Balanced Training		0.7462 (305.93)	0.9232 (621.85)	8.9		0.6533 (274.12)	0.6854 (554.35)	7.6
Co-Training+RAM		0.9123 (445.05)	0.7753 (674.14)	24.0		0.7503 (382.04)	0.5839 (574.91)	19.7
Our Approach		0.9194 (503.91)	0.9832 (719.08)	24.0		0.7762 (450.71)	0.7013 (615.05)	19.7
RAM	PAMAP2	0.7305 (184.25)	0.7359 (222.01)	27.8	MARS	0.7444 (126.83)	0.7148 (163.22)	21.5
Co-Training		0.6042 (212.38)	0.6837 (146.11)	10.3		0.6538 (118.22)	0.6104 (143.27)	8.2
Pattern-Balanced Training		0.6453 (471.48)	0.8629 (640.47)	10.3		0.737 (304.32)	0.7972 (606.97)	8.2
Co-Training+RAM		0.8085 (513.21)	0.7043 (683.27)	27.8		0.8233 (434.29)	0.6376 (644.72)	21.5
Our Approach		0.8125 (623.83)	0.9027 (753.98)	27.8		0.8325 (517.84)	0.8312 (697.19)	21.5

to overfitting of models. One-class classification is supposed to perform well on imbalanced data sets, but the performance is not that outstanding. Although the method is not affected by the class distribution, the insufficient data of those reduced classes still influence the training process. On the contrary, ensemble LSTM only has satisfactory results in $S1$, but it shows relatively better robustness to imbalanced data.

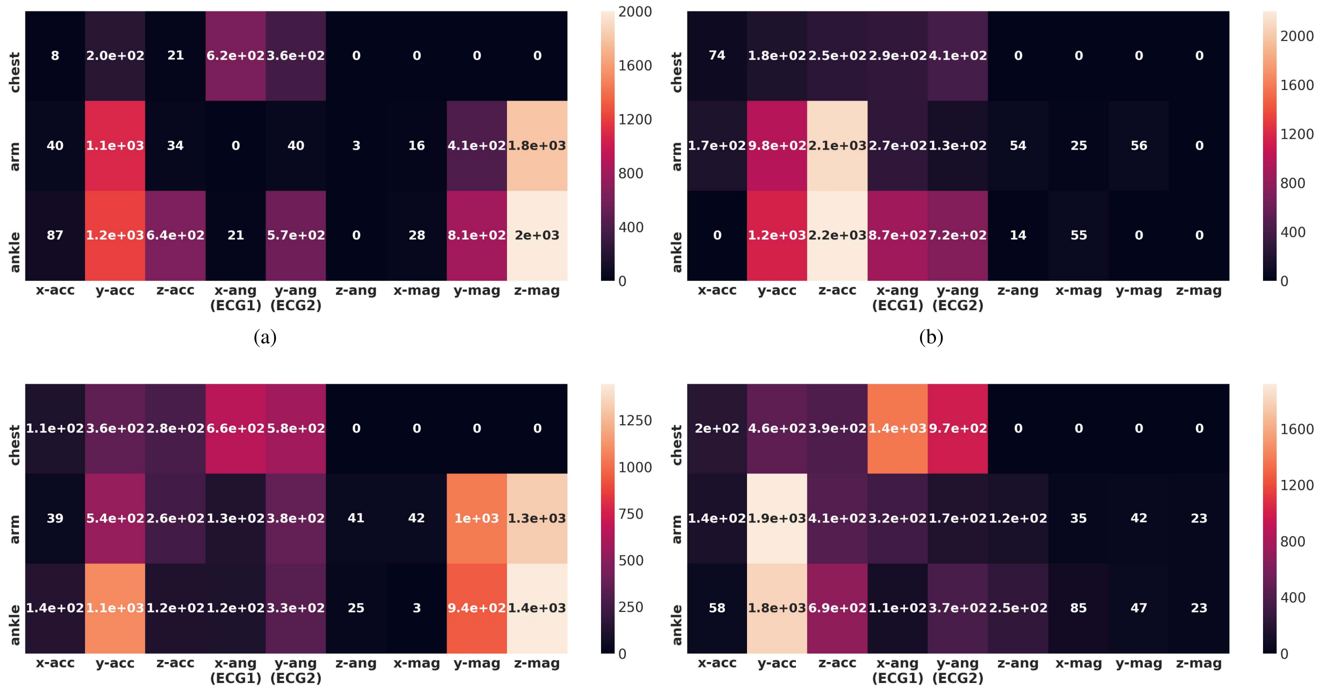
C. Ablation Study

We examine the effectiveness of the proposed components in our method in this section. Table IV presents the accuracy, training time, and test time of the ablation models on 2000 balanced data and the F1 scores on 5000 imbalanced training data. The imbalanced ratio is as S2 shown in Table II. To examine the contributions of cotraining framework and the pattern balanced training framework without the influence of RAM, we list the performance of these frameworks with regular CNNs. It can be seen that RAM-based methods need a longer time for the test as RAM has more parameters. There is a considerable increase in the training time when RAM is trained with cotraining since labeling is conducted in several training rounds. Pattern balanced training further increases the training time as it includes complex data processing. With respect to the performance, it can be observed that RAM in both situations obtains good performance. In the imbalanced situation, pattern-balanced training significantly improves the performance than regular cotraining since it mitigates the class imbalance. In our experiments, cotraining combined with RAM is not that effective on imbalanced data because the imbalance severely influences the F1 scores of classification. In a balanced situation, pattern-balanced training also makes an improvement. The reason for this is that the balanced labeled data may become imbalanced during training rounds, which is avoided by the pattern-balanced training. We can also observe that RAM combined with regular cotraining considerably improves the performance with its outstanding voting accuracy. Based on the observations, our method, composed of these components, is comparable with the state-of-the-art approaches.

D. Visualization of Selected Glimpses

The attention model extracts salient parts of the input sensory data for recognition, which makes the model explainable. In this section, we present visualized glimpses in recognizing standing, going upstairs, lying, and running on MHEALTH. The subjects wear sensors on their chest, dominant arms, and ankles, each sensor collecting multiple modalities. The available modalities collected from arms and ankles include three-axis acceleration, three-axis angular velocity, and three-axis magnetism, and from chests, the data set only includes three-axis acceleration and two ECG signals.

Fig. 4 shows the glimpse heatmaps of four activities. The glimpses are selected by a well-trained attention model for 10 000 times on input data that represent standing, going upstairs, lying, and running. The training is on 5000 labeled data and 15 000 unlabeled data in a semisupervised fashion. We observe that when recognizing a specific activity, the model does focus on only a part of modalities.



(a)

(b)

(c)

(d)

Fig. 4. Visualization of the selected glimpses on MHEALTH. Three rows represent modalities collected from chests, arms, and ankles, respectively. Each column denotes one modality. Acc, Ang, and Magn denote acceleration, angular velocity, and magnetism, respectively. Note that chests only contains three-axis acceleration and two ECG signals. The values in the grids represent the frequency with which this modality is selected. Lighter colors denote higher frequency. (a) Standing. (b) Going Upstairs. (c) Lying. (d) Running.

For example, magnetism (orientation) in standing and lying is selected as one of the most active features. The fact is that it is easy to distinguish between standing and lying with people's orientation. Another example is that the most distinguishing characteristic of going upstairs is "up". Therefore, z-axis acceleration is specifically selected by agents for going upstairs. Also, identifying running involves acceleration, ECG, and arm swing, which conforms to the experimental evidence as well. The model also selects several other features with lower frequencies, which avoids losing effective information.

V. Conclusion

This paper presents an integrated semisupervised activity recognition system based on multimodal wearable sensor data and addresses a rarely explored problem, i.e., semisupervised learning for imbalanced HAR. We first propose a cotraining framework that balances the latent patterns of activity data, and then deploy recurrent convolutional attention models as classifiers to exploit unlabeled samples. Comprehensive experiments conducted on four data sets validate the robustness and reliability of the proposed method.

References

[1] K. Chen, L. Yao, D. Zhang, X. Chang, G. Long, and S. Wang, "Distributionally robust semi-supervised learning for people-centric sensing," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 72–79.

[2] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, p. 33, 2014.

[3] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 1112–1123.

[4] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc.-Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 3995–4001.

[5] L. Yao, F. Nie, Q. Z. Sheng, T. Gu, X. Li, and S. Wang, "Learning from less for better: Semi-supervised activity recognition via shared structure discovery," in *Proc. ACM Ubicomp*, 2016, pp. 13–24.

[6] M. Stikic, K. Van Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," in *Proc. IEEE Wearable Comput. (ISWC)*, Sep./Oct. 2008, pp. 81–88.

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

[8] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee, "Activity recognition based on semi-supervised learning," in *Proc. 13th IEEE Int. Conf. Embedded Real-Time Comput. Syst. Appl. (RTCSA)*, Aug. 2007, pp. 469–475.

[9] Z.-H. Zhou, D.-C. Zhan, and Q. Yang, "Semi-supervised learning with very few labeled training examples," in *Proc. AAAI*, 2007, pp. 675–680.

[10] Y. Guan and T. Plötz, "Ensembles of deep LSTM learners for activity recognition using wearables," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, 2017, Art. no. 11.

[11] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification," in *Proc.-Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 22, no. 3, 2011, p. 1826.

[12] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[13] M. Stikic, D. Larlus, and B. Schiele, "Multi-graph based semi-supervised learning for activity recognition," in *Proc. IEEE Int. Symp. Wearable Comput. (ISWC)*, Sep. 2009, pp. 85–92.

[14] N. Y. Hammerla, J. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plötz, "PD disease state assessment in naturalistic environments using deep learning," in *Proc. AAAI*, 2015, pp. 1742–1748.

[15] M. Zeng, T. Yu, X. Wang, L. T. Nguyen, O. J. Mengshoel, and I. Lane, "Semi-supervised convolutional neural networks for human activity recognition," in *Proc. IEEE Big Data*, Dec. 2017, pp. 522–529.

[16] P. Juszczak and R. P. W. Duin, "Uncertainty sampling methods for one-class classifiers," in *Proc. Workshop Learn. Imbalanced Data Sets II (ICML)*, 2003, pp. 81–88.

[17] M. Frasca, A. Bertoni, M. Re, and G. Valentini, "A neural network algorithm for semi-supervised node label learning from unbalanced data," *Neural Netw.*, vol. 43, pp. 84–98, Jul. 2013.

[18] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.

[19] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1243–1251.

[20] K. Chen *et al.*, "Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 3016–3021.

[21] K. Chen, L. Yao, D. Zhang, B. Guo, and Z. Yu, "Multi-agent attentional activity recognition," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 4031–4038.

[22] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[23] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[24] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 499–508.

[25] N. J. Butko and J. R. Movellan, "I-POMDP: An infomax model of eye movement," in *Proc. 7th IEEE Int. Conf. Develop. Learn. (ICDL)*, Aug. 2008, pp. 139–144.

[26] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, 1992.

[27] V. Radu *et al.*, "Multimodal deep learning for activity and context recognition," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, 2018, p. 157.

[28] Y. Cheng, X. Zhao, R. Cai, Z. Li, K. Huang, and Y. Rui, "Semi-supervised multimodal deep learning for RGB-D object recognition," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 3345–3351.

[29] D.-D. Chen, W. Wang, W. Gao, and Z.-H. Zhou, "Tri-net for semi-supervised deep learning," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, Stockholm, Sweden, 2018, pp. 2014–2020.

[30] O. Banos *et al.*, "mHealthDroid: A novel framework for agile development of mobile health applications," in *Proc. Int. Workshop Ambient Assist. Living*. Cham, Switzerland: Springer, 2014, pp. 91–98.

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

[32] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. ESANN*, 2013, pp. 437–442.

Kaixuan Chen (S'03) is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of New South Wales (UNSW), Sydney, NSW, Australia.

Her current research interests include data mining, machine learning and their applications to Internet of Things, human activity recognition, and brain–computer interface.

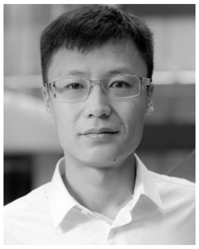
Lina Yao (M'11) is currently a Senior Lecturer with the School of Computer Science and Engineering, University of New South Wales (UNSW), Sydney, NSW, Australia.

Her current research interests include data mining and machine learning with applications to Internet of Things, information filtering and recommending, human activity recognition, and brain–computer interface.

Ms. Yao is a member of the ACM.

Dalin Zhang is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia.

His current research interests include data mining and machine learning with applications to human activity recognition, pervasive computing, and brain–computer interface.



Xianzhi Wang (M'15) is currently a Lecturer with the School of Software, University Technology of Sydney, Ultimo, NSW, Australia.

His current research interests include data management and service-oriented computing.

He is a member of the ACM.

Xiaojun Chang received the Ph.D. degree from the Center for Quantum Computation and Intelligent Systems (QCIS), University of Technology Sydney, Ultimo, NSW, Australia.

He is currently a Lecturer with Monash University, Clayton VIC, Australia. He is also a Post-Doctoral Associate with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interests include machine learning, data mining, and computer vision.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Professor with Northwestern Polytechnical University, Xian, China. His current research interests include machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.