A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach Using Multihead Convolutional Attention

Haoxi Zhang, Zhiwen Xiao, Juan Wang, Fei Li, and Edward Szczerbicki

***Abstract*—Together with the fast advancement of the Internet of Things (IoT), smart healthcare applications and systems are equipped with increasingly more wearable sensors and mobile devices. These sensors are used not only to collect data but also, and more importantly, to assist in daily activity tracking and analyzing of their users. Various human activity recognition (HAR) approaches are used to enhance such tracking. Most of the existing HAR methods depend on exploratory case-based shallow feature learning architectures, which struggle with correct activity recognition when put into real-life practice. To tackle this problem, we propose a novel approach that utilizes the convolutional neural networks (CNNs) and the attention mechanism for HAR. In the presented method, the activity recognition accuracy is improved by incorporating attention into multihead CNNs for better feature extraction and selection. Proof of concept experiments are conducted on a publicly available data set from wireless sensor data mining (WISDM) lab. The results demonstrate a higher accuracy of our proposed approach in comparison with the current methods.**

***Index Terms*—Attention mechanism, deep learning, human activity recognition (HAR), Internet of Things (IoT).**

I. Introduction

OVER the last decade, the concept of Internet of Things (IoT) has developed with an astounding pace [1], [2]. IoT's aptitude to integrate traditional networks, wearable sensors, and networked objects are the main causes for such fast development [1], [3], [4]. For example, body sensor networks (BSNs) [5] have emerged as one of the most effective IoT technologies enabling novel human-centered applications and new e-Health methods by combining wireless sensor and sensor networks placed in the human body, on the body surface, or around the body. Together with the technology of cloud computing, BSNs data can be further processed by clouds to allow not only powerful and flexible data storage and analysis but also to enable large-scale BSN applications, such as fitness, behavior surveillance, health care, and human activity recognition HAR [6], [7]. HAR as a novel application for implementing e-Health approaches has drawn considerable attention [8]. HAR attempts to recognize our daily activities that are important for numerous purposes, such as fitness tracking, home automation, motion mode detection, smart hospitals, mobility and transportation aged care, etc., which have a significant bearing on our personal well being.

Based on the embedded types of sensing modes, the HAR techniques can be divided into three groups: 1) radio-based HAR; 2) camera-based HAR; and 3) wearable-device-based HAR. The first group categorizes different human activities through the variations of wireless signal intensity [9]. Radio-frequency network system processing signal strength information for the joint purpose of human localization and detection proposed by Kianoush *et al.* [10] is a worthy example of radio-based HAR. The second group of the HAR techniques uses the computer vision technology to classify various human activities. For example, Liu *et al.* [11] introduced a class of short-term memory network for activity recognition by using a global context memory cell.

The third group applies built-in sensors, such as accelerometer, magnetometer, barometer, or gyroscope, to collect and classify the types of human activity-related information and data [12].

In this article, we propose a new wearable-device-based HAR architecture where the inputs are multichanneled time-series readings received from a set of inertial sensors of wearable devices, and the outputs are predefined classes of human activities. Typically, any HAR system includes data collection, preprocessing, segmentation, feature extraction, feature selection, modeling, and classification activities. The original data from sensors are processed to remove random noise and are assigned classes. Such preprocessed data are arranged into sequences. With the use of sequences, a number of features are obtained and selected to train the classifier. Subsequently, the activities can be classified by feeding sensor data into the classifier. The classification performance depends to much extend on the training features that are selected. In other words, extracting and selecting effective features play a key role in the success of properly identifying activities. This is a critical and extremely challenging task. Various time-series analysis techniques are often employed

Manuscript received May 29, 2019; revised September 1, 2019; accepted October 21, 2019. Date of publication October 25, 2019; date of current version February 11, 2020. This work was supported by the Sichuan Science and Technology Program under Grant 2019YFH0185. *(Corresponding author: Haoxi Zhang.)*

H. Zhang, Z. Xiao, J. Wang, and F. Li are with the School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China (e-mail: haoxi@cuit.edu.cn; xiao1994zw@163.com; wangjuan@cuit.edu.cn; lifei@cuit.edu.cn).

E. Szczerbicki is with the Faculty of Management and Economics, Department of Management, Gdansk University of Technology, 80-233 Gdansk, Poland (e-mail: edward.szczerbicki@zie.pg.gda.pl).

Digital Object Identifier 10.1109/JIOT.2019.2949715

to address this challenge. Techniques, such as symbolic representation [13], basis transform coding (e.g., signals with Fourier transform and wavelet transform) [14], and statistics of raw data (e.g., mean and variance of time sequences) [15], are often used in HAR. These techniques are heuristic and not task dependent [16]. Additionally, they are not robust [10], and extracting more training features does not necessarily improve the classification performance, but significantly increases the computational cost [12]. Furthermore, for typical HAR tasks, there are additional challenges, such as intraclass variability, interclass similarity, the NULL-class dominance, and complexness and diversity of physical activities [15], [16]. Recently, there is increasing interest in extending deep learning [17] approaches for other domains. Driven by the success of deep learning, researchers have borrowed ideas from image recognition [18], [19] to tackle the HAR problem [12], [16]. For example, Ronao and Cho [20] proposed a deep convolutional neural network (CNN) to perform efficient and effective HAR using smartphone sensors by exploiting the inherent characteristics of activities and 1-D time-series signals. Although these current deep-learning-based studies have shown advantages over previously existing methods, they might still achieve suboptimal performance. Therefore, there is a pressing need to develop an approach that can effectively extract and select features that can be used to train the classifier to effectively identify activities. Addressing this need, we propose a deep-learning-based approach for HAR, which employs the CNNs [18], and the attention mechanism introduced in [19]. CNN is used in the feature extraction process. The attention mechanism supports feature selection. The motivation behind our approach is based on the evidence that CNNs have proven extremely effective in extracting informative representations of data [17]. Attention mechanism, on the other hand, enables the presented approach to ignore the irrelevant features and to focus on a subset of pertinent features ensuring more accurate activity recognition. Based on the current state-of-the-art literature search in this field, this is the first work that applies to HAR multihead CNNs integrated with attention.

Compared with the existing HAR methods, the main advantage of our approach is a significant enhancement in activity recognition accuracy through improvement in both feature extraction and selection procedures. Unlike sheer CNN methods, such as presented in [16], the offered technique employs multihead convolution, which notably increases the variety of learned features. Also, our method exploits the power of attention mechanism for more effective feature selection. As shown by the experimental results presented in this article, the combination of the multihead convolution and attention achieves better recognition performance (recognition rate of 95.4% as measured by F-measure) than other well-known methods.

The rest of this article is organized as follows. Section II reviews the related literature. Section III presents the proposed HAR architecture, including data preprocessing, segmentation, and the relevant model. The experiments and results are illustrated and discussed in Section IV. Finally, in Section V, the conclusions are drawn.

II. Related Work

A. Multihead Convolutional Neural Networks

CNNs are constructed to process data that are presented in the form of manifold arrays. They are used to perform feature extraction and mapping of data [17]. There are four central sections in CNNs that take advantage of the intrinsic properties of natural signals: local connections, shared weights, pooling mechanism, and multilayer network structure [17]. All of these sections institute a well defined process that supports the extraction and mapping required for human activity identification as described in [17] and [18]. For the sake of completeness, it should be added here that CNN is a class of deep, feedforward artificial neural networks [16]. A supervised deep learning technique was the first computer-generated pattern recognizer to achieve human-competitive performance on certain recognition-related tasks. Moreover, when compared with the traditional feedforward networks, CNNs perform with much fewer connections, and so they are easier to train.

CNNs contain great promise to recognize patterns of HAR's signals. Computation units in the lower network layers attain the local basic features of activity signals, and computation units in the higher network layers extract the patterns of different activities at a higher level representation. The multihead CNNs [22] simply multiply this pattern extraction ability as the standard CNNs can be considered as one-head CNNs. With multiple heads, a CNN can have different filter banks and different processing layers in each head. For example, we can have a number of size 3×3 filters in head-1, and another number of filters sized 7×3 in head-2. If necessary, we may even choose whether to have dropout or pooling layers for a certain head. By using multiple heads, a CNN is equipped with the unique ability to allocate different feature learning polices to different components of the input signals, which is a promising facet for feature extraction in multichannel time-series signals received from the wearable sensors.

B. Attention

The attention mechanism can be described as mapping a query and a set of key–value pairs to an output, where the importance of each specific part of the input is computed as a weight according to its relativity to the output [23]. In other words, this procedure can help to assign a relevance score to elements in the input and to ignore the noisy parts [24]. Attention mechanisms have been successfully applied in a number of application domains, enhancing and improving object detection [25], image caption generation [26], speech recognition [27], machine translation [28], and question answering [29].

Instead of performing a single-attention function, multihead attention allows the model to jointly attend to information from different representation subspaces at different positions. Multihead attention mechanism has been reported with higher effectiveness in producing attention representation [21]. In our approach, the HAR's signals received from the wearable sensors are multichannel time-series data, which gives the multihead mechanism great potential to learn the relevance and importance of each piece of features produced by

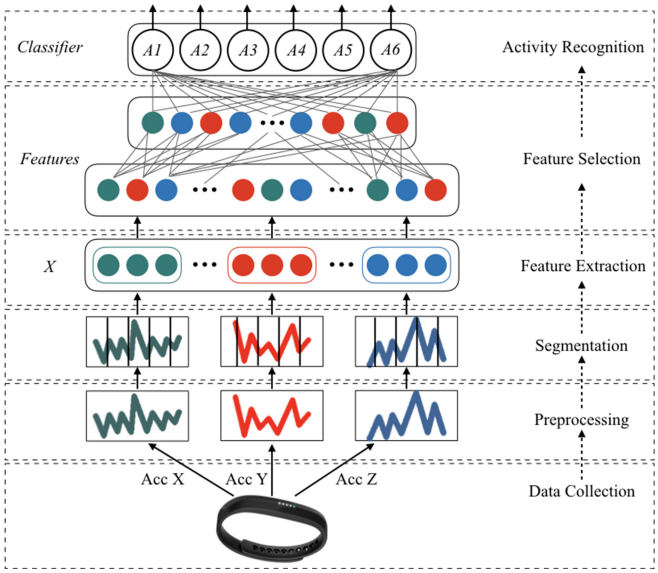


Fig. 1. Architecture of the proposed system.

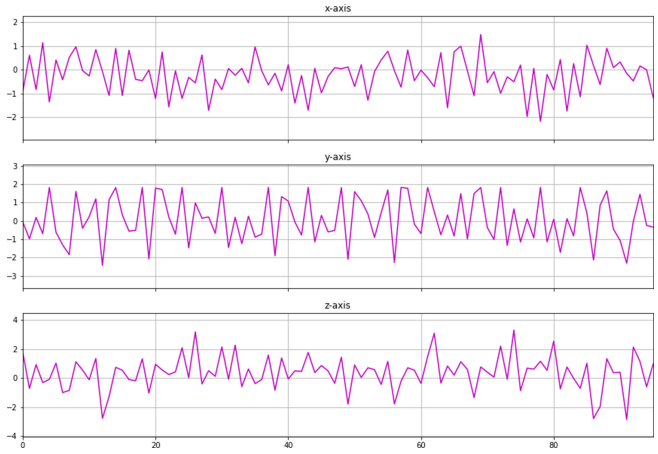


Fig. 2. Sample segmented sequence of the accelerometer sensor readings.

multihead CNNs, and eventually pick out the important ones for each activity, and thus improving the accuracy of activity recognition.

Motivated by the above promising findings, in this article, we propose to apply the attention mechanism to support the task of object detection. Instead of simply assembling attention module without any supervision, the attention block we propose is regularly nurtured with the instance segmentation annotations as the supervised input information.

III. Proposed Scheme

A. Architecture Overview

The architecture of the proposed system for HAR with multihead attention is shown in Fig. 1. It consists of data collection, preprocessing, segmentation, feature extraction, feature selection, and activity recognition steps. In the data collection stage, signals as well as their respective timestamps of built-in sensors, such as accelerometer, magnetometer, barometer, or gyroscope, are gathered and saved. It should be noted that in our case, only one accelerometer sensor is used. In the remainder of this section, we first introduce the human activities of interest, and then preprocessing step and segmentation procedure, followed by a detailed presentation of feature extraction and selection processes.

B. Human Activities

Human activities can be categorized into different classes [12], [30], including daily activities (shopping, using computer, sleeping, going to work, attending a meeting, etc.) health-related activities (e.g., falls, rehabilitation, following routines, and prescriptions), exercise (e.g., cycling and playing soccer), locomotion (*Walking, Running, Standing*, etc.), and so on. In this article, we focus on the locomotion activities and use the publicly available data sets wireless sensor data mining (WISDM) [31] in our experiments.

WISDM is the data set created by the Wireless Sensor Data Mining Lab based on smartphone accelerometer sensors under natural state. The data samples are recorded from performing six types of daily life activities, namely, *Walking, Jogging, Upstairs, Downstairs, Sitting,* and *Standing*.

C. Preprocessing

To be able to competently combine data from different sensors for activity recognition, sensor signals collected at the previous stage need to be aligned according to their respective timestamps. However, it is common that inertial sensors produce noisy data, which causes inaccurate readings and makes it difficult to catch the activity features reliably. Moreover, the sampling rate for the same sensor is not always stable [12]. Generally, in order to reduce the effect of noisy data and attain stabilization within a fixed-time window, one can use the filtering techniques, such as Kalman filter, low-pass filter, wavelet filter, etc., and utilize the spline interpolation to generate samples with a fixed sampling time interval [32]. Hence, how to tackle noisy data and signal instability lead the preprocessing to a crucial step that can dramatically affect the final performance of the model.

Nevertheless, this article skips data collection and preprocessing steps on purpose because it is meant to improve the HAR by providing better feature extraction and selection. Therefore, in this article, we use the publicly available as well as widely used data sets WISDM to examine the performance of our method rather than using private and elaborately preprocessed data set.

D. Segmentation

In activity recognition, a single point of data cannot provide the semantic information of a movement type, just like a single pixel in an image cannot give the meaning of the whole image content. As a result, the collected time-series signals need to be segmented into sequences according to a certain window size so that the information of activities can be carried in and eventually be used by recognition algorithms to classify the activities. Basically, the window size can be either fixed or

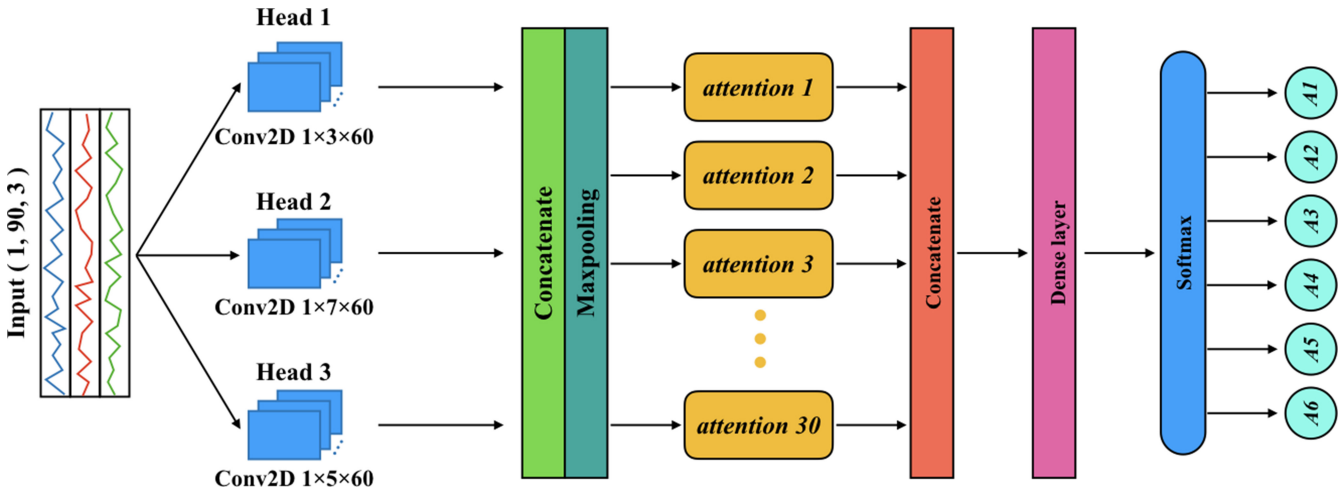


Fig. 3. Schematic of our proposed multihead convolutional attention approach. Given an input vector, we first use a 3-head CNN to extract its activity-related features. Then, effective ones are selected from these features using the multihead attention mechanism and passed to a dense layer. Finally, the softmax is used to give output for the activity category prediction.

adaptive. For example, Ma *et al.* [33] introduced a smart segmentation approach to recognition of human activities with adaptive time duration and achieved better activity recognition accuracy in both static and dynamic conditions. However, the fixed-size window methods are more computation efficient and friendly to end-to-end learning. Thus, we use a fixed-size window strategy in our proposed approach, even though the adaptive window methods might enhance the recognition accuracy. Besides, we are also curious whether our proposed approach is sensitive to the window sizes.

Specifically, one sequence is composed of three time series of accelerometer readings $\{S_i^{\text{acc}_x}, S_i^{\text{acc}_y}, S_i^{\text{acc}_z}\}$. Each series of readings corresponds to the data received from one of the three axes of an accelerometer, as shown in Fig. 2. These sequences are created using a sliding window as follows:

$$S_i^{\text{acc}_x} = \left[\text{acc}_t^x, \text{acc}_{t+1}^x, \ldots, \text{acc}_{t+K-1}^x\right] \quad (1)$$

$$S_i^{\text{acc}_y} = \left[\text{acc}_t^y, \text{acc}_{t+1}^y, \ldots, \text{acc}_{t+K-1}^y\right] \quad (2)$$

$$S_i^{\text{acc}_z} = \left[\text{acc}_t^z, \text{acc}_{t+1}^z, \ldots, \text{acc}_{t+K-1}^z\right] \quad (3)$$

where K is the size of the sliding window. K values of 48, 64, and 90 are used in our experiments, the results of which are presented in Section IV.

E. Feature Extraction

Feature extraction is a crucial procedure in HAR. In this part, we present the proposed feature extraction method based on multihead CNNs.

We start with the notation used in the multihead CNN. Let S_i represent the input vector at time i (4), which is a 2-D matrix containing $K \times D$ sensor readings, where K is the size of the sliding window, and D represents the dimension of the sensor readings, i.e., 90×3 in our case. For the training data, the true label of the matrix instance is determined by the most frequently occurred label of K raw samples

$$S_i = \left[S_i^{\text{acc}_x}, S_i^{\text{acc}_y}, S_i^{\text{acc}_z}\right]. \quad (4)$$

In order to extract various features, a 3-head CNN is designed to process the input vector, as shown in Fig. 3. In the convolution layers, the previous layer's feature maps are convolved with a set of convolutional kernels (to be learned in the training process). The output of the convolution operators enhanced by a bias (to be learned) is put through the activation function to form the feature map for the next layer. Formally, the jth feature map at the ith layer of cth head of the multihead CNN is also a matrix, and the value at the xth row is denoted as $v_{ij}^{x,c}$, and it is given by

$$v_{ij}^{x,c} = f_{\text{ReLU}}\left(f_{\text{conv2d}}^c\left(v_{i-1}^{x+p}\right)\right) \qquad \forall c = 1, 2, 3 \quad (5)$$

where f_{ReLU} is the activation function that replaces all negative values in the feature map by zero, and f_{conv2d}^c is the convolution function of the cth head in our multihead CNN, as presented in

$$f_{\text{conv2d}}^c\left(v_{i-1}^{x+p}\right) = b_{ij} + \sum_m \sum_{p=0}^{n_i^c - 1} w_{ijm}^{p,c} v_{(i-1)m}^{x+p,c} \quad (6)$$

where b_{ij} is the bias for this particular feature map, m is the index of the feature maps at the (i-1)th layer connected to the current feature map, $w_{ijm}^{p,c}$ is the value at the position p of the convolutional kernel, and n_i^c is the length of the kernel at the ith layer of cth head of the multihead CNN. After the feature extraction is followed, this procedure provides a number of various features, and sends them to the next step.

F. Feature Selection

The HAR problem cannot be solved effectively by simply using features that are extracted. In our approach, we propose that extracted features are to be further selected and categorized according to their contribution to activity recognition. The attention mechanism is employed to calculate this contribution. Attention mechanism maps a query and a set of key–value pairs to an output, where the importance of each specific part of the input is computed as a weight according

Algorithm 1 Multihead Convolutional Attention Method

Input: labeled activity recognition dataset: $D = \{X_i, Y_i\}$
Output: activity label y_i of the test data
// *Initialization:*
Initialize the parameters θ
Normalize the dataset
Segment the normalized data into sequences:
training dataset: $D_{\text{train}} = \{S_i^{train}, Y_i^{train}\}$
validation dataset: $D_{\text{val}} = \{S_i^{val}, Y_i^{val}\}$
testing dataset: $D_{\text{test}} = \{S_i^{test}, Y_i^{test}\}$
// *Training on training and validation datasets*
for episode=1, M **do**
  for n = 1, N **do**
    get the input vector $S_i \in D_{\text{train}}$
    feedforward the S_i and get the output y_i
    compute $L = -\sum_i (y_i log\ \tilde{y}_i + (1 - y_i) log\ (1 - \tilde{y}_i))$
    perform a gradient decent step on $(L \mid \theta)$
    if (n%20 == 0) **then**
      validate the model using D_{val}
    end if
  end for
end for
// *Testing*
Use the trained network to predict the labels y_i of the testing dataset D_{test}

to its relativity to the output

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q is the query matrix, and V and K are the matrices of keys and values. To exploit features from different representation subspaces extracted via different convolution channels, multihead attention is further utilized to perform parallel attention function h times. The multihead attention is calculated as

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)\end{aligned} \tag{8}$$

where W_i^Q, W_i^K, and W_i^V are the weight matrices in parallel attentions with dimensions d_k/h, d_k/h, and d_v/h, respectively. W^O is the output weight matrix with dimension d_O. Finally, to train the model, we minimize the cross-entropy loss

$$\text{loss} = -\sum_i (y_i \log \tilde{y}_i + (1 - y_i) \log(1 - \tilde{y}_i)) \tag{9}$$

where y_i is the correct activity label, while $\tilde{y}_i$ is the prediction label given by our method for the input vector S_i.

In this article, we employ $h = 30$ parallel attention heads, and for each head, we use $d_k/h = d_v/h = 128$. By combining multihead CNN with multihead attention, our approach can effectively extract and select features, progressively enhancing the activity-relevant representation learning for HAR. The pseudocode for the multihead convolutional attention method is summarized in Algorithm 1.

IV. Experiments and Results

In this section, we first introduce the details of the data set used in experiments. Then, we introduce the basic descriptions of involved evaluation measurements. Finally, we show the experimental results of our proposed approach.

TABLE I
Percentage of Samples of Each Class in WISDM

Activities	Instances	Proportion
Walking (A1)	424,400	38.6%
Jogging (A2)	342,177	31.2%
Upstairs (A3)	122,869	11.2%
Downstairs (A4)	100,427	9.1%
Sitting (A5)	59,939	5.5%
Standing (A6)	48,395	4.4%
Total	1,098,207	100%

TABLE II
Confusion Matrix of Classification Results

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

A. Data Set Description

As introduced in Section III, we use the WISDM data set in our experiments. There are six different activities in this data set, namely, *Walking* (A1), *Jogging* (A2), *Upstairs* (A3), *Downstairs* (A4), *Sitting* (A5), and *Standing* (A6), and 1 098 207 examples in total. The detailed information of this data set is listed in Table I.

B. Evaluation Measurements

In this article, *Acc*, *Pre*, *Rec*, and F_1 are employed to evaluate the final classification performance of our proposed approach in HAR.

Acc is the overall accuracy for all classes calculated as

$$\text{Acc} = \frac{1}{M}\sum_{i=1}^{M}\frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \tag{10}$$

where TP is the number of true positive instances, TN is the number of true negative instances, FP is the number of false positive instances, and FN is the number of false negative instances, as presented in Table II. Pre is the precision of correctly classified positive instances to the total number of instances classified as positive

$$\text{Pre} = \frac{1}{M}\sum_{i=1}^{M}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}. \tag{11}$$

Rec is the recall of correctly identified positive instances to the total number of actual positive instances

$$\text{Rec} = \frac{1}{M}\sum_{i=1}^{M}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \tag{12}$$

TABLE III
TEST ACCURACY COMPARISON OF DIFFERENT SEGMENT SIZES WITH DIFFERENT CNN-BASED METHODS

K	Class	1D-CNN	2D-CNN	Multi-head 2D-CNN	Multi-head Convolutional Attention
48	A1	92.0	74.0	96.1	96.0
	A2	94.2	96.0	91.2	93.0
	A3	90.0	89.0	96.7	98.0
	A4	92.0	97.0	89.0	98.0
	A5	88.6	90.0	97.8	97.0
	A6	86.0	91.0	87.4	93.0
	Overall	90.5	89.5	93.0	**95.8**
64	A1	93.5	70.0	96.1	96.0
	A2	91.0	98.0	91.2	94.0
	A3	94.0	89.0	97.0	97.0
	A4	93.5	98.0	89.0	98.0
	A5	89.5	90.0	98.0	99.0
	A6	86.0	93.0	87.8	92.0
	Overall	91.3	89.7	93.2	**96.0**
90	A1	86.5	82.0	94.0	98.0
	A2	94.0	95.0	91.2	97.5
	A3	94.3	89.0	98.0	99.0
	A4	93.0	97.0	92.0	97.0
	A5	91.8	86.0	92.6	88.0
	A6	89.2	85.0	90.0	99.0
	Overall	91.5	89.0	93.0	**96.4**

F_1 is a key evaluation measure of classification performance, which considers both the precision and the recall of the test. In our experiments, it is calculated as

$$F_1 = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}}. \tag{13}$$

C. Experiment Analysis and Performance Comparison

In this section, we report and discuss experiments and results from a number of HAR approaches under different validation methods.

First, we analyze the influence of sliding window size K on the accuracy of activity recognition, as well as the classification performance of different CNN-based methods, which is shown in Table III. As it can be seen from Table III, the size of the sliding window (i.e., K) has an impact on the recognition accuracy. For example, when K value is 90, the accuracy of *Sitting* (A5) is declined significantly. The main reason for this might be that longer sequences will contain more signals (better chance from other activities) that might be more interesting to our model than plain signals from a still sitting activity does, hence such signals catch the "attention" of our model and mislead the results. We obtain the highest accuracy when K is set to 90. In our experiments, we also test the performance of different CNN-based methods, namely, 1-D CNN (1D-CNN), 2-D CNN (2D-CNN), multihead 2D-CNN, and our proposed approach, i.e., the multihead convolutional attention. By adding multihead attention to the multihead 2D-CNN, our approach achieves better performance with 96.4% as measured by testing accuracy, which demonstrates that the multihead attention mechanism does provide the feature learning in HAR with a noticeable enhancement.

TABLE IV
CONFUSION MATRIX OF TESTING OF THE PROPOSED APPROACH

		Prediction					
		A1	A2	A3	A4	A5	A6
Actual	A1	**601**	17	0	0	22	32
	A2	4	**2220**	0	0	0	38
	A3	2	0	**421**	3	2	1
	A4	0	2	0	**330**	0	1
	A5	9	34	0	0	**626**	139
	A6	2	2	0	0	3	**2811**

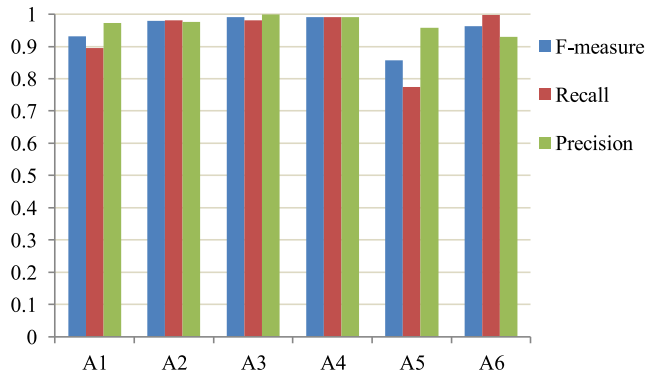


Fig. 4. Recognition performance for each activity.

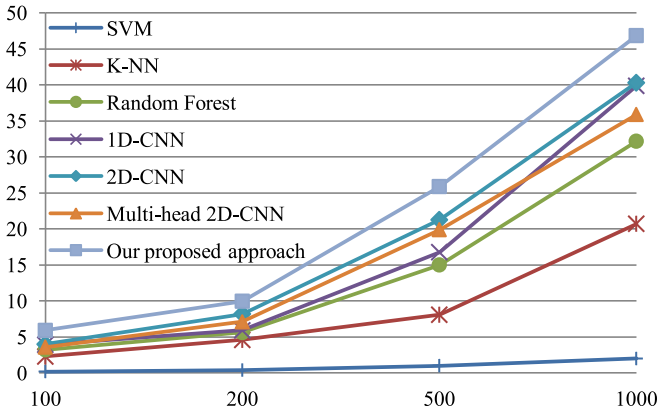


Fig. 5. Computational complexity comparison by test set size.

Second, we analyze the results of our proposed approach for recognizing each activity, which are summarized in Table IV and Fig. 4. There are 139 out of 808 segments of *Sitting* (A5) are misidentified as *Standing* (A6). The main reason for this might be that both sitting and standing are still; hence the accelerometer readings are similar. Meanwhile, there are 38 out of 2262 segments of *Jogging* (A2) and 32 out of 672 segments of *Walking* (A1) incorrectly classified as *Standing* (A6). The reason is that standing activity has significantly more samples than other activities, which could make the recognition results biased toward standing.

Third, from an IoT perspective, the computation cost is also an important factor that we need to take into account. Therefore, we compare the computational complexity with

TABLE V
COMPUTATIONAL COMPLEXITY COMPARISON OF THE PROPOSED APPROACH WITH SOME OTHER METHODS

Test set size	Parameters	Methods	With GPUs	Only CPU
100	-	SVM	0.012ms	0.21ms
	-	KNN	0.13ms	2.29ms
	-	Random Forest	0.22ms	3.22ms
	772,960	1D-CNN	0.256ms	3.92ms
	1,124,334	2D-CNN	0.297ms	4.00ms
	715,334	Multi-head 2D-CNN	0.250ms	3.599ms
	2,771,270	Our proposed approach	0.378ms	5.92ms
200	-	SVM	0.021ms	0.41ms
	-	KNN	0.27ms	4.59ms
	-	Random Forest	0.43ms	5.62ms
	772,960	1D-CNN	0.502ms	5.92ms
	1,124,334	2D-CNN	0.697ms	8.20ms
	715,334	Multi-head 2D-CNN	0.45ms	7.10ms
	2,771,270	Our proposed approach	0.746ms	9.97ms
500	-	SVM	0.06ms	1.02ms
	-	KNN	0.63ms	8.09ms
	-	Random Forest	1.20ms	15.01ms
	772,960	1D-CNN	1.12ms	16.78ms
	1,124,334	2D-CNN	1.51ms	21.27ms
	715,334	Multi-head 2D-CNN	1.01ms	19.9ms
	2,771,270	Our proposed approach	1.68ms	25.89ms
1000	-	SVM	0.115ms	2.01ms
	-	KNN	2.067ms	20.72ms
	-	Random Forest	2.104ms	32.19ms
	772,960	1D-CNN	2.46ms	39.87ms
	1,124,334	2D-CNN	2.84ms	40.3ms
	715,334	Multi-head 2D-CNN	2.5ms	35.9ms
	2,771,270	Our proposed approach	3.78ms	46.85ms

TABLE VI
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH WITH OTHER EXISTING METHODS

Authors	Methods	F-measure
Lu et al. [5] (2018)	SVM	0.802
	BAGGING	0.813
	KNN	0.752
	ST+Time	0.936
Kwapisz et al. [27] (2011)	J48	0.851
	Logistic Regression	0.781
	Multi-Perceptron	0.917
Zdravevski et al. [28] (2017)	SVM *on mHealth dataset*	0.934
Gu et al. [9] (2018)	Stacked Denoising Autoencoders on their own dataset	0.940
Our proposed approach	Multi-head Convolutional Attention	**0.954**

some other existing methods. In our experiments, four different sizes (100, 200, 500, 1000) of test sets generated from the WISDM data set are used ($K = 90$). All experiments are carried out on a computer with two Nvidia 1070Ti 8G GPUs and one Intel i7-8700 CPU, which are used for "With GPUs" and "Only CPU" tests, respectively. The results are shown in Table V and Fig. 5. As can be seen from Table V, our proposed approach is most computationally expensive, which gives 100 predictions in around 5.92 ms by using CPU. Fig. 5 shows how to test set size impacts on the computational complexity of different methods with CPU (and the results of tests with GPUs are very similar). The SVM is the most computationally efficient and stable method compared to the other approaches.

Finally, we use F-measure as the metric to compare the performance of our proposed approach with other existing approaches presented in the literature (Table VI). Except for the methods presented by Gu *et al.* [12] and Zdravevski *et al.* [34] which are using different data sets, the remaining seven methods listed in Table VI use the WISDM data set. In all instances, the proposed activity recognition approach performs better than the current benchmark classifiers achieving a high classification rate of 95.4% as measured by F-measure.

V. CONCLUSION

In this article, we presented a novel, IoT-perseptive, approach to HAR based on accelerometer sensor. The proposed architecture integrates multihead convolution neural networks with attention mechanism for better feature extraction and selection. The proposed approach does not require manual feature engineering. It automatically learns the effective features for activity classification. The experimental results show that the proposed approach can achieve very high recognition rate of 96.4% as measured by testing accuracy and 95.4% as measured by F-measure. Bearing IoT concept in mind, HAR accuracy and effectiveness turn out to be of utmost importance for a number of obvious reasons. The accuracy results of the proposed system demonstrate its relevance to perform activity identification with very high confidence and could make inroads into numerous future applications.

Our future work will involve exploring ways to reduce the complexity of our proposed approach and make it more practical. We believe that a detailed study of elaborately handcrafted features and automatically learned features needs to be performed first. Then, we plan to distill the prior knowledge encoded in these features and introduce such knowledge into neural networks to enhance the model with long-term dependencies that are hard to learn with a limited data set. This will help us to reduce the model's complexity because long-term dependencies are not modeled with network connections, thus it scales down our model as well as the expected training data set. In addition, it is also part of our future work to create a new data set. By collecting more data from a diverse group of people using the accelerator sensor on a wrist and including equal samples of different activities will enrich the training data set and further improve the final results, especially when our system is applied to real-world users' wrist-worn devices.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

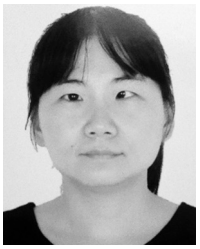
[2] K. Ashton, "That 'Internet of Things' thing," *RFID J.*, vol. 22, no. 7, pp. 97–114, Jun. 2009.
[3] G. Kortuem, F. Kawsar, V. Sundramoorthy, and D. Fitton, "Smart objects as building blocks for the Internet of Things," *IEEE Internet Comput.*, vol. 14, no. 1, pp. 44–51, Dec. 2009.
[4] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 414–454, Jan. 2014.
[5] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," *Inf. Fusion*, vol. 22, pp. 50–70, Mar. 2015.
[6] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "BodyCloud: A SaaS approach for community body sensor networks," *Future Gener. Comput. Syst.*, vol. 35, pp. 62–79, Jun. 2014.
[7] G. Fortino *et al.*, "Cloud-based activity-aaService cyber–physical framework for human activity monitoring in mobility," *Future Gener. Comput. Syst.*, vol. 75, pp. 158–171, Oct. 2017.
[8] W. Lu, F. Fan, J. Chu, P. Jing, and Y. Su, "Wearable computing for Internet of Things: A discriminant approach for human activity recognition," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2749–2759, Apr. 2019.
[9] S. Wang and G. Zhou, "A review on radio based activity recognition," *Digit. Commun. Netw.*, vol. 1, no. 1, pp. 20–29, Feb. 2015.
[10] S. Kianoush, S. Savazzi, F. Vicentini, V. Rampa, and M. Giussani, "Device-free RF human body fall detection and localization in industrial workplaces," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 351–362, Apr. 2017.
[11] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.
[12] F. Gu, K. Khoshelham, S. Valaee, J. Shang, and R. Zhang, "Locomotion activity recognition using stacked denoising autoencoders," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2085–2093, Jun. 2018.
[13] J. Lin, E. J. Keogh, S. Lonardi, and B. Y.-C. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proc. 8th ACM SIGMOD Workshop Res. Issues Data Mining Knowl. Disc.*, Jun. 2003, pp. 2–11.
[14] T. Huynh and B. Schiele, "Analyzing features for activity recognition," in *Proc. Joint Conf. Smart Objects Ambient Intell. Innov. Context Aware Services Usages Technol.*, Oct. 2005, pp. 159–163.
[15] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surveys*, vol. 46, no. 3, p. 33, Jan. 2014.
[16] 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. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, Buenos Aires, Argentina, Jul. 2015, pp. 3995–4001.
[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[18] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.
[19] M. M. Hassan, G. R. Alam, Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Inf. Fusion*, vol. 51, pp. 10–18, Nov. 2019.
[20] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.
[21] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[22] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 94–98, Jan. 2019.
[23] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. NIPS*, 2010, pp. 1243–1251.
[24] H. Du and J. Qian, "Hierarchical gated convolutional networks with multi-head attention for text classification," in *Proc. 5th Int. Conf. Syst. Informat. (ICSAI)*, Nanjing, China, 2018, pp. 1170–1175.
[25] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention CoupleNet: Fully convolutional attention coupling network for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2019.
[26] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[27] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.
[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473
[29] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 21–29.
[30] O. D. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *J. Bionanosci.*, vol. 3, no. 2, pp. 145–171, 2013.
[31] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explor. Newslett.*, vol. 12, no. 2, pp. 74–82, 2011.
[32] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," *IEEE Access*, vol. 5, pp. 3095–3110, 2017.
[33] C. Ma, W. Li, J. Cao, J. Du, Q. Li, and R. Gravina, "Adaptive sliding window based activity recognition for assisted livings," *Inf. Fusion*, vol. 53, pp. 55–65, Jan. 2020.
[34] E. Zdravevski *et al.*, "Improving activity recognition accuracy in ambient assisted living systems by automated feature engineering," *IEEE Access*, vol. 5, pp. 1–17, 2017.

Haoxi Zhang received the Ph.D. degree in knowledge engineering from the University of Newcastle, Callaghan, NSW, Australia, in 2013, and the master's degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China.

He is an Associate Professor at the Chengdu University of Information Technology, Chengdu. He has published over 30 reputed journals and conference papers. His current research interests include experience-oriented intelligent systems, knowledge engineering, Internet of Things, and deep learning.

Zhiwen Xiao was born in Ya'an city, China, in 1994. He is currently pursuing the undergraduation degree at the Internet of Things Engineering, Chengdu University of Information Technology, Chengdu, China.

His current research interests include deep learning and computer vision.



Juan Wang was born in Chengdu, China, in 1981. He received the B.S. degree in computer science, the M.S. degree in computer architecture, and the Ph.D degree in information security from the University of Electronics and Technology of China, Chengdu, in 2003, 2006, and 2010, respectively.

She was a Visiting Scholar with the University of North Carolina at Charlotte, Charlotte, NC, USA, from 2007 to 2008 studied on network flow analysis. She is currently an Associate Professor with the School of Cybersecurity, Chengdu University of Information Technology, Chengdu. She participated in multiple national or provincial scientific research project. Her current research interests include network security, Internet of Things security, especially the intelligent vehicle security, and their application.

Fei Li received the B.E. degree in Internet of Things and the M.E. degree in computer science automatic control from the University of Science and Technology of Chengdu, Chengdu, China, in 1988 and 1993, respectively.

He is currently a Professor and the Dean of School of Cybersecurity, Chengdu University of Information Technology, Chengdu. His current research interests include field of network and information system security, vehicle intelligence and security, Internet of Things technology and applications, and mobile Internet applications.

Edward Szczerbicki received the D.Sc. degree in information science from the Szczecin University of Technology, Szczecin, Poland, in 1993.

He had very extensive experience in the area of intelligent systems development over an uninterrupted 40 year period, 25 years of which he spent in top systems research centers in the United States, U.K., Germany, and Australia. In this area, he contributed to the understanding of information and knowledge management in systems operating in environments characterized by informational uncertainties. He has published close to 350 refereed papers with over 2000 citations over the last 20 years. His academic experience includes ongoing positions with Gdansk University of Technology, Gdansk, Poland; Strathclyde University, Glasgow, Scotland; the University of Iowa, Iowa City, IA, USA; the University of California at Berkeley, Berkeley, CA, USA; and the University of Newcastle, Callaghan, NSW, Australia. He has given numerous invited presentations and addresses at universities in Europe, USA, and at international conferences. He received the Title of Professor of information science for his international published contributions in 2006.

Prof. Szczerbicki serves as a Board Member of Knowledge Engineering Systems, and a Member of Berkeley Initiative in Soft Computing Special Interest Group on Intelligent Manufacturing. He is a Member of the Editorial Board/Associated Editor for eight international journals. He chaired/co-chaired and acted as a committee member for a number of international conferences.