Post-print of: Zhang H., Xiao Z., Wang J., Li F., Szczerbicki E.: A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach Using Multi-Head Convolutional Attention IEEE Internet of Things Journal Vol. 7, iss. 2 (2019), pp.1072 - 1080 DOI: 10.1109/JIOT.2019.2949715

A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach

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

***Abstract*—**

Together with 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 control 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 straggle 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 multi-head convolutional neural networks for better feature extraction and selection. Proof of concept experiments are conducted on a publicly available dataset from Wireless Sensor Data Mining (WISDM) laboratory. The results demonstrate higher accuracy of our proposed approach in comparison with the current methods.

***Index Terms*—Internet of Things, human activity recognition, deep learning, attention mechanism**

I. Introduction

OVER the last decade, the concept of Internet of Things (IoT) has developed with 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]. Together with the progress of IoT, HAR as a novel application for implementing smart healthcare approaches has drawn significant attention [5]. 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, HAR techniques can be divided into three groups: radio-based HAR, camera-based HAR, and wearable device-based HAR. The first group categorizes different human activities through the variations of wireless signal intensity [6]. Radio-frequency network system processing signal strength information for the joint purpose of human localization and detection proposed by Kianoush et al. [7] is a worthy example of radio-based HAR. The second group of HAR techniques uses the computer vision technology to classify various human activities. For example, Liu et al. [8] 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 (Gu et al. [9].

In this paper, we propose a new wearable device-based HAR architecture where inputs are multichanneled time series readings received from a set of inertial sensors of wearable devices, and outputs are predefined classes of human activities. Typically, any HAR system includes data collection, preprocessing, segmentation, feature extraction, feature selection, modelling, 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, 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 plays 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 to address this challenge. Techniques like symbolic representation [10], basis transform coding (e.g. signals with Fourier transform and wavelet transform) [11], and statistics of raw data (e.g, mean and variance of time sequences) [12], are often used in HAR. These techniques are heuristic and not task-dependent [13]. Additionally, they are not robust [7], and extracting more training features does not necessarily improve the classification performance, but significantly increases the computational cost [9]. 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 [12][13]. The current HAR approaches are not able to address all the above challenges. Therefore there is a

Manuscript received May xxx, 2019; revised xxxxxx; accepted xxxxxx. Date of publication xxxxxx; date of current version xxxxxx. 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 Gdansk University of Technology, Gdansk, Poland (e-mail: edward.szczerbicki@zie.pg.gda.pl).

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 convolutional neural networks (CNNs) [14], and the attention mechanism introduced in [15]. 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 [16]. 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 multi-head convolution neural networks integrated with attention.

Compared with existing HAR architectures, the main advantage of our approach is a significant enhancement in activity recognition accuracy through improvement in both critical procedures, extraction and selection. Unlike sheer CNN methods, such as presented in [13], the offered technique employs multi-head convolution, which notably increases the variety of leant features. Also, our method exploits the power of attention mechanism for more effective feature selection. As shown by experimental results presented in the paper, the combination of the multi-head 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 the paper 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. Multi-head Convolutional Neural Networks

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 [16]. There are four central sections in CNNs that take advantage of the intrinsic properties of natural signals: local connections, shared weights, pooling mechanism, and multi-layer network structure [16]. All of these sections institute a well defined process that supports the extraction and mapping required for human activity identification as described in [14][16]. For the sake of completeness, it should be added here that CNN is a class of deep, feed-forward artificial neural networks [13]. Supervised deep learning technique was the first computer-generated pattern recognizer to achieve human-competitive performance on certain recognition related tasks. Moreover, when compared to traditional feed-forward 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 multi-head CNNs [17] 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 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 [18]. In other words, this procedure can help to assign a relevance score to elements in the input and to ignore the noisy parts [19]. Attention mechanisms have been successfully applied in a number of application domains, enhancing and improving object detection [20], image caption generation [21], speech recognition [22], machine translation [23], and question answering [24].

Instead of performing a single attention function, multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. Multi-head attention mechanism has been reported with higher effectiveness in producing attention representation [25]. In our approach, the HAR's signals received from wearable sensors are multichannel time series data, which gives the multi-head mechanism great potential to learn the relevance and importance of each piece of features produced by multi-head convolutional neural networks, 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 paper 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 multi-head attention is shown in Fig. 1. It consists of data collection, preprocessing, segmentation, feature extraction, feature selection, and activity recognition steps. It should be noted that in the data collection step, only one accelerometer sensor is used. Therefore, the data are composed of single accelerometer sensor's signals with respective timestamps. In the reminder of this Section, we first introduce human activities of interest, then the segmentation procedure, followed by detailed presentation of feature extraction and selection processes.

B. Human Activities

Human activities can be categorized into different classes [9] [26], including daily activities (shopping, using computer, sleeping, going to work, and attending a meeting, etc), health related activities (for example falls, rehabilitation, following routines and prescriptions), exercise (e.g., cycling, playing soccer), locomotion (walking, running, standing, etc), and so on. In this paper, we focus on the locomotion activities, and use the publicly available datasets WISDM [27] in our experiments.

WISDM is the dataset created by the Wireless Sensor Data Mining (WISDM) Lab based on smart-phone 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. 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. Therefore, we segment the sensor readings into sequences according to a certain time frame. Specifically, one sequence is composed of three time series of accelerometer readings $\{ S_i^{acc_x}, S_i^{acc_y}, S_i^{acc_z} \}$. 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:

$$S_i^{acc_x} = [acc_t^x,\ acc_{t+1}^x, \dots, acc_{t+K-1}^x] \quad (1)$$

$$S_i^{acc_y} = [acc_t^y,\ acc_{t+1}^y, \dots, acc_{t+K-1}^y] \quad (2)$$

$$S_i^{acc_z} = [acc_t^z,\ acc_{t+1}^z, \dots, acc_{t+K-1}^z] \quad (3)$$

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.

D. Feature Extraction

Feature extraction is a crucial procedure in HAR. In this part, we present the proposed feature extraction method based on multi-head convolutional neural networks.

We start with the notation used in the multi-head CNN. Let S_i represent the input vector at time i (Eq. 4), which is a

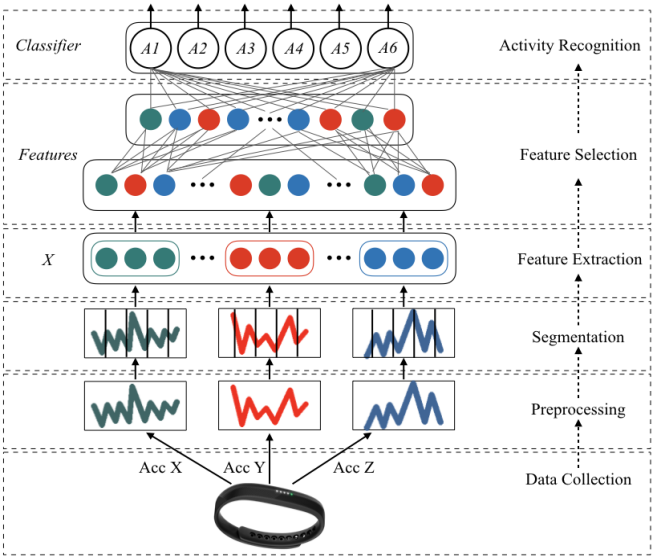

Fig. 1.
Architecture of the proposed system.

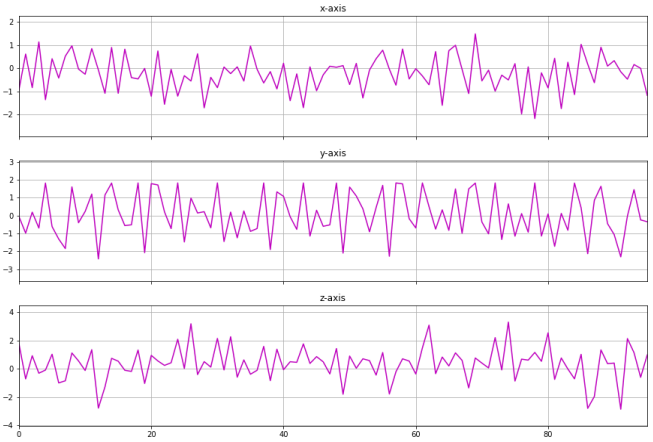

Fig. 2.
A sample segmented sequence of the accelerometer sensor readings.

two-dimensional matrix containing $K \times D$ sensor readings where K is

$$S_i = [S_i^{acc_x}, S_i^{acc_y}, S_i^{acc_z}] \quad (4)$$

the size of the sliding window, and D represents the dimension of the sensor readings, i.e. 90×3 in our case. For training data, the true label of the matrix instance is determined by the most-frequently occurred label of K raw samples.

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 j^{th} feature map at the i^{th} layer of c^{th} head of the multi-head

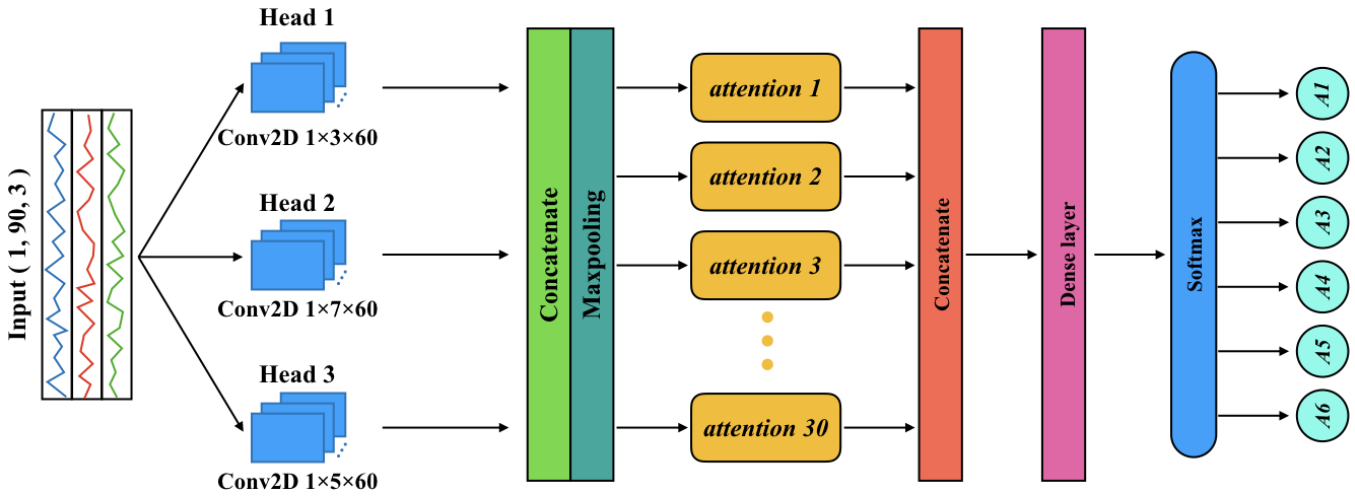

Fig. 3. Schematic diagram of our proposed Multi-head Convolutional Attention approach. Given an input vector, we first use a 3-head CNN to extract its activity-related features. Then, effective ones are selectedfrom these features using multi-head attention mechanism and passed to a dense layer. Finally,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

CNN is also a matrix, and the value at the x^{th} row is denoted as $v_{ij}^{x,c}$, and it is given by Eq (5):

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

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 c^{th} head in our multi-head CNN, as presented in Eq. (6):

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

where b_{ij} is the bias for this particular feature map, m is the index of the feature maps at the $(i\text{-}1)^{\text{th}}$ 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 i^{th} layer of c^{th} head of the multi-head CNN. After the feature extraction is followed, this procedure provides a number of various features, and sends them to the next step.

E. Feature Selection

The human activity recognition 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 to its relativity to the output:

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

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

$$MultiHead(Q,K,V) = Concat(\text{□}ead_1, \dots \text{□}ead_{\text{□}})W^O$$
$$\text{where } head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{8}$$

where W_i^Q, W_i^K, W_i^V are the weight matrices in parallel attentions with dimensions d_k/h, d_k/h, 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:

$$\text{loss} = -\sum_{i}\left(y_i log\ \tilde{y}_i + (1-y_i) log\ (1-\tilde{y}_i)\right) \tag{9}$$

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 work, we employ h = 30 parallel attention heads, and for each head we use $d_k/h = d_v/h$ = 128. By combining multi-head CNN with multi-head attention, our approach can effectively extract and select features, progressively enhancing the activity-relevant representation learning for HAR. The pseudo-code for the Multi-head Convolutional Attention method is summarized in Algorithm 1.

IV. Experiments and Results

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

A. Dataset Description

As introduced in Section III, we use the WISDM dataset in our experiments. There are six different activities in this dataset,

Alogrithm 1: Multi-head 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}

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%

namely *Walking* (A1), *Jogging* (A2), *Upstairs* (A3), *Downstairs* (A4), *Sitting* (A5), and *Standing* (A6). The detailed information of this dataset is listed in Table I.

B. Evaluation Measurements

In this paper, *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:

$$Acc = \frac{1}{M}\sum_{i=1}^{M}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \qquad (10)$$

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:

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)

$$Pre = \frac{1}{M}\sum_{i=1}^{M}\frac{TP_i}{TP_i + FP_i} \ . \qquad (11)$$

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

$$Rec = \frac{1}{M}\sum_{i=1}^{M}\frac{TP_i}{TP_i + FN_i} \ . \qquad (12)$$

The 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:

$$F_1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \ . \qquad (13)$$

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

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**

the accuracy of activity recognition, as well as classification performance of different CNN-based methods, which is shown in Table III.

As it can be seen from Table III, the size of sliding window (i.e. *K*) has an impact on the recognition accuracy. 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-dimentional CNN (1D-CNN), 2-dimentinal CNN (2D-CNN), multi-head 2D-CNN, and our proposed approach, i.e. the multi-head convolutional attention. By adding multi-head attention to the multi-head 2D-CNN, our approach achieves better performance with 96.4% as measured by testing accuracy, which demonstrates that the multi-head attention mechanism does provide the feature learning in HAR with a noticeable enhancement.

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.

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 V). Except for the methods presented by Zdravevski et al. [28] and Gu et al. [9] which are using different datasets, the remaining seven methods listed in Table V use WISDM dataset. 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. With the IoT concept in mind

V. Conculsion

In this paper, we present a novel, IoT-perseptive, approach to human activity recognition based on accelerometer sensor. The proposed architecture integrates multi-head 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.

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**

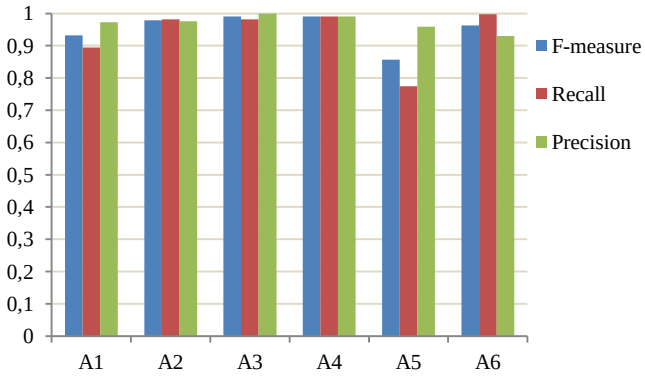

Fig. 4. Recognition performance for each activity.

TABLE V
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**

References

[1] Atzori L, Iera A, Morabito G. The internet of things: A survey. Computer networks. 2010 Oct 28;54(15):2787-2805.
[2] Ashton K. That 'internet of things' thing. RFID journal. 2009 Jun 22;22(7):97-114.
[3] Kortuem G, Kawsar F, Sundramoorthy V, Fitton D. Smart objects as building blocks for the internet of things. IEEE Internet Computing. 2009 Dec 1;14(1):44-51.
[4] Perera C, Zaslavsky A, Christen P, Georgakopoulos D. Context aware computing for the internet of things: A survey. IEEE communications surveys & tutorials. 2014 Jan 1;16(1):414-454.

[5] Lu W, Fan F, Chu J, Jing P, Su Y. Wearable Computing for Internet of Things: A Discriminant Approach for Human Activity Recognition. In Proc., IEEE Internet of Things Journal. 2018.
[6] Wang S, Zhou G. A review on radio based activity recognition. Digital Communications and Networks. 2015 Feb 1;1(1):20-29.
[7] Kianoush S, Savazzi S, Vicentini F, Rampa V, Giussani M. Device-free RF human body fall detection and localization in industrial workplaces. IEEE Internet of Things Journal. 2017 Apr; 4(2):351-362.
[8] Liu J, Wang G, Duan LY, Abdiyeva K, Kot AC. Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Transactions on Image Processing. 2018 Apr; 27(4):1586-1599.
[9] Gu F, Khoshelham K, Valaee S, Shang J, Zhang R. Locomotion activity recognition using stacked denoising autoencoders. IEEE Internet of Things Journal. 2018 Jun;5(3):2085-2093.
[10] Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. InProceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery 2003 Jun 13 (pp. 2-11). ACM.
[11] Huynh T, Schiele B. Analyzing features for activity recognition. InProceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies 2005 Oct 12 (pp. 159-163). ACM.
[12] Bulling A, Blanke U, Schiele B. A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR). 2014 Jan 1;46(3):33.
[13] Yang, J.B., Nguyen, M.N., San, P.P., Li, X.L., Krishnaswamy, S.. Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina, 25–31 July 2015; pp. 3995–4001.
[14] LeCun, Y. et al. Handwritten digit recognition with a back-propagation network. In Proc. Advances in Neural Information Processing Systems 396–404 (1990).
[15] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. InAdvances in neural information processing systems 2017, pp. 5998-6008.
[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[17] S. Ö. Arık, H. Jun and G. Diamos, "Fast Spectrogram Inversion Using Multi-Head Convolutional Neural Networks," in IEEE Signal Processing Letters, vol. 26, no. 1, pp. 94-98, Jan. 2019.
[18] Larochelle, Hugo and Hinton, Geoffrey E. Learning to combine foveal glimpses with a third-order boltzmann machine. In NIPS, pp. 1243–1251, 2010.
[19] H. Du and J. Qian, "Hierarchical Gated Convolutional Networks with Multi-Head Attention for Text Classification," 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, 2018, pp. 1170-1175.
[20] Zhu, Y., Zhao, C., Guo, H., Wang, J., Xu, Z., & Lu, H.. Attention couplenet: fully convolutional attention coupling network for object detection. IEEE Transactions on Image Processing, VOL. 28, NO. 1, pp. 113-126, JAN. 2019.
[21] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057, 2015.
[22] 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., pp. 577–585, 2015.
[23] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
[24] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 21–29, 2016.
[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[26] O. D. İncel, 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.
[27] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, pp. 74–82, 2011.
[28] E. Zdravevski, P. Lameski, V. Trajkovik, A. Kulakov, I. Chorbev, R. Goleva, N. Pombo, and N. Garcia, "Improving activity recognition accuracy in ambient assisted living systems by automated feature engineering," IEEE Access, pp.1-17, 2017.

Haoxi Zhang is an Associate Professor from the Chengdu University of Information Technology, Chengdu, China. He received his Ph.D. degree in Knowledge Engineering from the University of Newcastle in 2013, and the master's degree in Software Engineering from the University of Electronic Science and Technology of China. His research interests focus on experience-oriented intelligent systems, knowledge engineering, Internet of Things, and Deep Learning. He has published more than 30 reputed journal and conference papers.

Zhiwen Xiao was born in Ya'an city, Sichuan province, China, in 1994. He is an undergraduate student studying the Internet of Things Engineering at the Chengdu University of Information Technology. He is going to pursue a Ph.D. degree in computer science after his undergraduate study. His research interests are deep learning and computer vision.

Juan Wang was born in Chengdu, Sichuan province of China in 1981 and received her B.S. degree of computer science in 2003, and the M.S. degree of Computer Architecture and Ph.D degree of Information Security from University of Electronics and Technology of China (UESTC) in 2006 and 2010. And being a visiting scholar at University of North Carolina at Charlotte(UNCC) from 2007.9 to 2008.9 studied on network flow analysis. Now, she is an associate professor in School of Cybersecurity of Chengdu University of Information Technology(CUIT).She participated in multiple national or provincial scientific research project, and now her research interests include network security, IoT(Internet Of Things) security, especially the intelligent vehicle security, and their application.

Fei Li received the B.E. degree in Internet of things from University of

Science and Technology of Chengdu, Chengdu, Sichuan, China, in 1988 and M.E. degrees in computer science automatic control from the same university, in 1993. He is currently a Professor and the Dean of School of Cybersecurity, Chengdu University of Information Technology, Chengdu, Sichuan, China. His research interests are in the field of network and information system security, vehicle intelligence and security, Internet of things technology and applications and mobile Internet applications.

Edward Szczerbicki has 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 centres in the USA, UK, 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 more than 2000 citations over the last 20 years. His D.Sc. degree (1993) and the Title of Professor (2006) were gained in the area of information science for his international published contributions.

Professor Szczerbicki serves as a Board Member of Knowledge Engineering Systems (KES), and a Member of Berkeley Initiative in Soft Computing Special Interest Group on Intelligent Manufacturing. He has given numerous invited presentations and addresses at universities in Europe, USA and at international conferences. 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. His academic experience includes ongoing positions with Gdansk University of Technology, Gdansk, Poland; Strathclyde University, Glasgow, Scotland; The University of Iowa, Iowa City, USA; University of California, Berkeley, USA; and The University of Newcastle, Newcastle Australia.