An Enhanced Human Activity Recognition Algorithm with Positional Attention

Chenyang Xu CHYOND.XU@GMAIL.COM
Intelligent Manufacturing Department, Wu Yi University, Jiangmen, Guangdong, 529020, China
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100090, China

Jianfei Shen SHENJIANFEI@ICT.AC.CN
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100090, China
Shandong Academy of Intelligent Computing Technology, Jinan, Shandong, 250102, China

Feiyi Fan FANFEIYI@ICT.AC.CN
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100090, China

Tian Qiu TIMEQIU@HOTMAIL.COM
Intelligent Manufacturing Department, Wu Yi University, Jiangmen, Guangdong, 529020, China

Zhihong Mao MZH_YU@126.COM *
Intelligent Manufacturing Department, Wu Yi University, Jiangmen, Guangdong, 529020, China

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Human activity recognition (HAR) attracts widespread attention from researchers recently, and deep learning is employed as a dominant paradigm of solving HAR problems. The previous techniques rely on domain knowledge or attention mechanism extract long-range dependency in temporal dimension and cross channel correlation in sensor's channel dimension. In this paper, a HAR model with positional attention (PA), termed as PA-HAR, is presented. To enhance the features in both sensor's channel and temporal dimensions, we propose to split the sensor signals into two 1D features to capture the long-range dependency along the temporal-axis and signal's cross-channel information along the sensor's channel-axis. Furthermore, we embed the features with positional information by encoding the generated features into pairs of temporal-aware and sensor's channel-aware attention maps and weighting the input feature maps. Extensive experiments based on five public datasets demonstrate that the proposed PA-HAR algorithm achieves a competitive performance in HAR related tasks compared with the state-of-the-art approaches.

Keywords: Human activity recognition (HAR), Positional Attention (PA), Wearable Device, Deep learning.

1. Introduction

Utilizing data collected from embedded sensor for the purpose of HAR is a highly active research area in ubiquitous computing. HAR system based on wearable devices detects various daily activities (such as jogging, walking), and more complicated activities like assembly line worker behaviors. Hence, HAR system has a wide application in mobile social networks, environmental monitoring, and health monitoring. For example, in health monitoring, HAR system is used to track individuals' diseases like Alzheimer, Parkinson, and

* Corresponding Author: Zhihong Mao

emergencies like children fall detection. More specifically, HAR system can help Alzheimer patients' family record patients' daily behaviors (such as sitting, walking, and running) and detect abnormal behaviors (such as falling) throughout the disease process, to prevent events that may endanger patients' health. We use the inertial measurement unit (IMU) in the wearable device to measure the values of the accelerometer, gyroscope, magnetometer. Then, the data is preprocessed, which needs to fill the missing values and resample the gyroscope, accelerometer, and magnetometer signal to adjust to a uniform sampling frequency. Finally, we concatenate the multiple channel signals. Then, we use sliding window technology to split the signal data of multiple channels to signal images. The existed approaches mostly split the sensor signal into fixed-size sequences and then classify the signal frame by activity type using various machine learning methods. Qian et al. (2018a) employ a distribution kernel embedding method to extract all orders of moments as statistical features. Semwal et al. (2021b) employ the extreme learning machine (ELM) for HAR. Yang et al. (2014) employ ELM for timeliness online sequential predicting. Such conventional machine learning methods depend significantly on handcrafted features, which are largely constrained by human domain knowledge.

An increasing number of researchers introduce deep neural networks (DNN) to solve HAR problems. DNN is constructed of various layers, such as convolutional neural networks (CNN) layer, recurrent neural networks (RNN) layer, and graph convolutional networks (GCN) layer, etc. It does not require domain expertise and significantly improves the detection accuracy over the conventional machine learning algorithms.

In previous studies, temporal-spatial DNN methods are often used to extract high-level information. Specifically, CNN is used to extract the temporal and channel characteristics of specific patterns in the data segment. Along this research line, researchers focus on techniques like combining channel and temporal attention (Gao et al. (2020)) or make use of shallow CNN with channel-selectivity (Huang et al. (2021b)). However, aforementioned methods can only extract the local temporal dependency and sensor's cross-channel relationship. Some researchers introduce RNN to solve the above problems. RNN realizes the re-modulation of historical signal by adding hidden layer and it is well-suited for exploiting the temporal relationships within a specific acvitity (Francisco and Daniel (2016)), while it introduces noise (irrelevant signal components and sensor modalities, etc.) in encoding process (Zeng et al. (2018)). Therefore, researchers equip RNN with temporal attention and sensor attention to capture the spatial-temporal interdependence of sensor signals (Dua et al. (2021)). Additionally, with the adoption of GCN (Huang et al. (2021a)), only a shallow network is needed to extract the interaction between different feature channels. But GCN only considers convolution feature channel relationship and ignores the sensor's cross-channel relationship as well as temporal interaction.

The state-of-the-art HAR methods are mostly combinations of aforementioned research lines, where 2D convolution and global pooling are used to extract temporal and cross-channel relationship simultaneously. However, modeling the long-range temporal dependency and cross-channel requires further investigation.

With the rise of attention mechanisms in computer vision area, some scholars combine attention mechanisms with convolutional neural networks. Squeeze-and-Excitement (SE) attention (Jie et al. (2017)) and Convolution Block Attention Module (CBAM) attention (Woo et al. (2018)) are most commonly adopted attention mechanism. SE attention mech-

anism includes two steps: squeeze and excitation, which are used for embedding global information and feature channel's adaptive recalibration, respectively. However, SE attention use 2D global pooling process to capture global information, which only consider the correlation between the feature channel, and will lose relationship in temporal dimension and sensor's dimension. CBAM attention mechanism employs 7×7 convolution kernel to capture the local cross sensor's channel information in spatial dimension and local time sequence dependency in temporal dimension. And then they establish correlation between the feature channel. However, they ignore the global long-range dependency and cross sensor's channel correlation. As self-attention is proposed, the Transformer model based on self-attention becomes popular. Transformer uses position embeding to establish correlation signal image patch and then extract local feature in signal image patch (Dirgová Luptáková et al. (2022)), which only considers the local association between signal image patch. While the global positional information is ignored. Hou et al. (2021) propose coordinate attention, which can dynamically increase the weights on both $H - direction$ and $W - direction$ in image area. Inspired by this concept, we propose the HAR method with PA mechanism.

In this paper, we present a novel and efficient HAR method with PA mechanism. Different from position encoding in transformer (Dirgová Luptáková et al. (2022)), the PA mechanism embeds global positional information into both temporal and sensor's channel dimensions to enable the network to extract long-range temporal dependency and sensor's cross-channel information simultaneously. It avoids loss of positional information due to 2D global pooling, as well as the increased computing cost introduced by enlarged network structure for modeling the long-range temporal dependency.

In PA mechanism, firstly, we employ two 1D global pooling processes to respectively aggregate the input features along the temporal-axis and sensor's channel-axis into two separate direction-aware feature maps. These two feature maps with embedded temporal-channel positional information are encoded individually into two attention maps, which capture long-range dependency along with temporal-axis and sensor's cross-channel information along with sensor's channel-axis, respectively. Secondly, we employ two 1×1 convolution processes to extract interaction information of two attention maps between each convolution feature channel. Thirdly, the input feature map is multiplied with both attention maps to enhance the areas of sharp signal changes. Finally, we combine PA mechanism with CNN for HAR to enhance the temporal-channel positional information.

The contributions of our paper are summarized as follows.

- We propose a novel positional attention mechanism for sensor-based HAR, which encourages to capture long-range dependency along with temporal-axis and introduce sensor channel-axis cross-channel relationship. Furthermore, we can extract the interaction between the convolution feature channel.

- We employ two baseline networks (Resnet and 3-layers CNN), and add the PA mechanism to baseline network. Only using sample 1×1 Conv and 1D global pooling processes, we get a greater performance than baseline. Therefore we demonstrate that the gained features are more diversified and discriminative.

- Extensive experiments on various benchmark datasets are carried out to illustrate the higher performance of our proposed PA-HAR.

The following is a description of the paper's structure. We summarize HAR's relevant work in Section II. The specifics of HAR with the PA method are presented in Section III. In Section IV, we present the public HAR dataset and the experiment design in detail and compare experiment results on a variety of levels and our discussion. We describe conclusions in Section V.

2. RELATED WORKS

This section gives a brief literature review of HAR and attention mechanism, including previous works on feature engineering approaches, deep learning approaches and attention mechanism.

2.1. Feature engineering approches

Feature engineering approches include Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), both of which are based on transformation coding (such as the wavelet and Fourier transforms) and the handcrafted signal features such as the median and moment order (like mean, variance, and skewness) (Figo et al. (2010)). Some machine learning methods, like support vector machines (SVM) (Bulling et al. (2011)) and random forests (RF) (Stisen et al. (2015)) use these features to predict what people are doing at a given time. Guo et al. (2018) propose an entropy-based hierarchical fusion model for HAR, consisting of a sensor fusion layer and a classifier fusion layer. With the rise of extreme learning machine (ELM) method, some researchers (Semwal et al. (2021a)) use ELM method to recognize human activity, and the result is better than SVM. Qian et al. (2018b) use the kernel mean embedding technique to automatically extracts all orders of moments as statistical features. Besides the statistical features, some methods also look at extra meta-information as structural features. For example, empirical cumulative distribution function (ECDF) method preserves the overall shape and spatial information of time series data (Hammerla et al. (2013)).

2.2. Deep learning approches

Deep learning approaches employ DNN to automaticly perform feature extraction and classification, which provide promising results for HAR. CNN is used to identify human activities and process the original sensor signal into 2D signal images. Previous study has shown that this signal image can capture important spatial attributes and is better than conventional machine learning methods (Hammerla et al. (2016)). Huang et al. (2021b) propose a channel selection method: firstly, the Expected Channel Damage Matrix (ECDM) (Jeong and Shin (2019)) is used to identify the contribution of each convolution feature channel, and the feature channel with a low contribution rate is recovered in the training stage, secondly, the channel with a high contribution rate is reassigned to the position of the recovered low contribution feature channel. Teng et al. (2020) propose a local loss method to avoid using global backpropagation over the whole network, which can make the weights be updated along with forward pass. Dirgová Luptáková et al. (2022) use transformer to classify activities. Wang et al. (2019) propose a CNN that employs the based on attention method to

detect weakly labeled sensor signal data. Gao et al. (2020) propose adding CBAM attention to CNN.

Meanwhile, RNN is presented and successfully used in HAR, which can capture long-term data in time series. Zeng et al. (2018) combine LSTM with two attention mechanisms for HAR: temporal attention and spatial attention. These mechanisms can be used to highlight significant parts of time series sensor data and sensor's channel. To ensure the continuity of time series signals, RNN must include regularization terms for temporal and sensor attention. Ma et al. (2019) propose AttnSense, a novel attention-based multimodal neural network model. AttnSense provides a framework for capturing the relationships between signal in both spatial and temporal dimensions by combining attention processes with Gated Recurrent Units (GRU) networks. This method shows advantages in prioritized sensor selection and improves comprehensibility.

Another extensively used member of the deep learning family is made use of a combination of CNN layers and RNN layers (Dua et al. (2021)). Qian et al. (2019) propose meaningful features of automatic learning, including statistical features, temporal features, and spatial features, in which the statistical module's goal is to discover sensor location correlations, the temporal module's goal is to discover temporal sequence relationships, the spatial module's goal is to discover sensor channel positional correlations.

In recent literature, several neural network-based GCN models is proposed. Huang et al. (2021a) propose a shallow CNN model that extracts cross convolution feature channel communication in the HAR area, where all convolution feature channels in the same layer interact comprehensively to capture more discriminative aspects of sensor data. One convolution feature channel can interact with all other channels via GCN to reduce redundant information accumulated across channels, which is more efficient for constructing lightweight deep models.

2.3. Attention mechanisms

Attention processes are advantageous for a variety of tasks in computer vision, including image classification and segmentation. SENet (Jie et al. (2017)) is one of the effective examples, which efficiently squeezes each 2D feature map to generate interdependencies among convolution feature channels. CBAM (Woo et al. (2018)) extends this idea by adding spatial attention through convolution process with a large kernal size. Later research, such as GALA (Linsley et al. (2018)) and AA (Bello et al. (2020)), build on this concept by employing other spatial attention processes or building improved attention blocks. Dual Attention Network (DAN) (Fu et al. (2020)) uses Non-Local based convolution feature channel attention and spatial attention for semantic segmentation. As self-attention is proposed, the transformer model based on self-attention becomes popular. Wu et al. (2021) propose vision transformer, which establish position embedding in image by divide signal image to image patch. And then, they extract features in every image patch. Finally, they establish a global association between the learned position information and local feature information. All the above strategies improve the performance in computer vision area by learning effective attention imformation.

3. METHODOLOGY

This section introduces the PA mechanism in detail. We denote multiple raw sensor signals to a predetermined window size as $S = \{S_1, S_2, \ldots, S_n\}$, where $S_i \in \mathbf{R}^{m \times n}$, S_i is signal image given to network, m is the length of the time series, and n denotes the channel dimension.

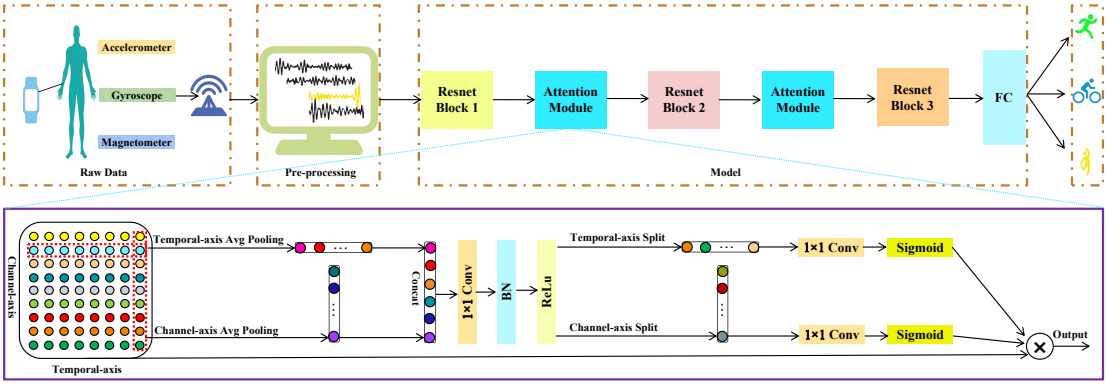


Figure 1: The Model of Positional Attention Based on Resnet.

3.1. Model Overview

HAR is challenging as it is affected by many factors, such as activity duration difference and participant's activity difference. Based on the above motivation, we design an end-to-end trainable neural network structure for HAR. The proposed network structure consists of two key operations to learn HAR related feature representations.

- Positional information embedding: This operation extracts long-range interaction through single dimension global pooling.
- Positional weights generation: This operation uses convolution and nonlinear operations to generate positional weights information and then multiplies it with original input.

We can develop a trainable model for the HAR problem by putting the above learned features together and establishing an unified architecture. The overall illustration of the proposed model is shown in Fig.1.

3.2. Positional Information Embedding

The PA mechanism encodes both sensor's cross-channel relationships and long-range dependencies. It is important to include the attention process in the HAR model, as suggested by Ma et al. (2019). Therefore, we use PA mechanism in HAR classfication. We discover that the global average pooling procedure builds connections inside the feature channel, which can increase the model's sensitivity to the feature channels information. For given input $X \in \mathbb{R}^{C \times H \times W}$, where the X is the signal feature map, the global average pooling step can

be formulated as follows:

$$P = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j) \tag{1}$$

where P is the output of the global average pooling. In Eqn.(1), the global pooling operation will loss the positional correlation information.

$$p^h(h) = \frac{1}{W} \sum_{0 \le i < W} X(h,i) \tag{2}$$

$$p^w(w) = \frac{1}{H} \sum_{0 \le j < H} X(j,w) \tag{3}$$

To encode positional correlation information along temporal and sensor's channel dimensions, we employ two directional pool kernals $(H,1)$ and $(1,W)$ depending on the input X. Where the $W-direction$ is responsible for the relationship information between sensor's channels, and the $H-direction$ captures temporal long-range dependency.

The aforementioned procedures are described by Eqn.(2) and Eqn.(3). These two improvements enable the attention mechanism to capture long-range dependencies along a single temporal-axis and maintain correlations among sensor's channel-axis, allowing the networks to find the items of interest more precisely.

3.3. Positional Weights Generation

Eqn.(2) and Eqn.(3), as previously stated, allow a global receptive field and exact positional information. To utilize the resultant expressive representations, we introduce the second design, called positional weights generation. Our design is based on the three standard listed below. Firstly, the new design should be structurally simple and computationally efficient. Secondly, it can fully utilize the acquired positional information to precisely highlight the areas of interest. Thirdly, it should be able to capture inter-channel connections well, which are shown to be important in previous research (Jie et al. (2017)).

Our positional weights generation method is as follows. Firstly, in order to normalize the enhanced features, we concatenate the outputs of Eqn.(4) and Eqn.(5). Secondly, we use the 1×1 convolution process and sigmoid activation function to induce nonlinearity.

$$mid = \delta\left(Conv\left(\left[p^h, p^w\right]\right)\right) \tag{4}$$

where δ is sigmoid activation function, $Conv$ is convolution function and $[\cdot,\cdot]$ is concatenation operation. Finally, we split the mid into two distinct tensors along the spatial dimension (mid^h and mid^w), then apply two 1×1 convolutions to revert to the input size. The formula is listed below:

$$out^h = \sigma\left(Conv\left(mid^h\right)\right) \tag{5}$$

$$out^w = \sigma\left(Conv\left(mid^w\right)\right) \tag{6}$$

where σ is the sigmoid activation function, $Conv$ is the 1×1 convolution operation. After that, the outputs out^h and out^w are enlarged and utilized as attention weights. Meanwhile, our PA mechanism output Y is expressed as follows:

$$Y(i,j) = X(i,j) \times out^h(i) \times out^w(j) \tag{7}$$

Unlike convolution feature channel attention, that examines the relevance of distinct channels, our PA mechanism incorporates spatial information encoding. The input tensor X receives attention in both horizontal and vertical directions simultaneously. The area of interest appears in the relevant row and column in each element of the two attention maps. This encoding process permits our PA mechanism to selecet the exact position of the area of interest, which aids the overall model's recognition.

3.4. Implementation

As this paper's aim is to research a more effective method for augmenting the convolutional features of HAR networks, we use two architectures (Resnet and 3-layers CNN) to show the benefits of PA mechanism over previous HAR methods. The Resnet only includes three sets of convolutional blocks, each of which consists of two convolutional layers of the same kernel size. In Resnet, we add the PA mechanism to the former two sets, and in 3-layers CNN, we add the PA mechanism to every layer. The Fig.1 shows the constructure of the Resnet with PA mechanism.

4. EXPERIMENT

4.1. Datasets

In order to evaluate the effectiveness of the proposed model, we conduct extensive experiments based on five public HAR datasets. The UCIHAR dataset (Anguita et al. (2012)), PAMAP2 dataset (Reiss and Stricker (2012)), UniMib-SHAR dataset (Micucci et al. (2017)), DSADS (Altun et al. (2010)), and MHEALTH dataset (Banos et al. (2014)) are employed as the five benchmark HAR datasets. The classification number, division proportion, and window size of the datasets are shown in Table 1, and the class description of five daatasets is shown in Table 2.

Table 1: Briefly Description of The Operation for The HAR Datasets.

Operation \ Dataset	UCIHAR	PAMAP2	UniMib-SHAR	DSADS	MHEALTH
Number of Classification	6	12	17	9	13
Ratio of Train-set	70%	80%	70%	70%	75%
Ratio of Test-set	30%	20%	30%	30%	25%
Sliding Window Size	128	512	151	*None*	100
Overlap Rates	50%	50%	50%	*None*	50

UCIHAR Dataset This dataset is collected with a group of 30 participants ranging in age from 19 to 48 years old. Each subject conducted six actions while wearing a smartphone (Samsung Galaxy S II) on their waist, including walking, going upstairs, walking downstairs,

Table 2: Class Description of UCIHAR, PAMAP2, UniMib-SHAR, DSADS and MHEALTH Datasets.

Dataset	The Category of Activity
UCIHAR	Walking, Walking_Upstairs, Walking_Downstairs, Sitting, Standing, Lying
PAMAP2	Lying, Sitting, Standing, Walking,Running, Cycling, Nordic Walking Descending Stairs, Vacuum Cleaning, Ironing, Rope Jumping, Ascending Stairs
UniMib-SHAR	StandingUpFS, StandingUpFL, Running, SittingDown, GoingUps, FallingBack, Syncope, Jumping, FallingLeft, GoingdownS, Walking, Falling with PS LyingDownS, FallingFrow, HittingObstacle, FallingRight, FallingBackSC
DSADS	Moving Around in an Elevator, Standing in an Elevator Still, Playing Basketball, Walking on a Treadmill 4 km/h (in Flat Position and in 15 Deg Inclined Position), Exercising on a Stepper, Exercising on a Cross Trainer, Descending Stairs, Moving Around in an Elevator, Walking in a Parking Lot
MHEALTH	Standing Still, Sitting and Relaxing, Lying Down, Walking, Climbing Stairs, Waist Bends Forward, Frontal Elevation of Arms,Knees Bending, Cycling, Jogging, Running, Jump front & back, NULL

lying, sitting, and standing. The accelerometer and gyroscope sensors data streams are sampled at a rate of 50 Hz to capture the subjects' actions. There are nine characteristics in the raw time series data (i.e., body acceleration, total acceleration, and gyroscope signals in all three directions).

PAMAP2 Dataset This dataset is collected from nine volunteers. The volunteers performed 12 mandatory different activities including walking, cycling, rope jumping, etc. Multiple sensors including chest sensor, wrist sensor and ankle sensor are applied to record the data. The 100 Hz sampling rate is downsampled to 33.3 Hz for further analysis. Sport intensity is estimated using a heart rate monitor with a sample rate of 9 Hz.

UniMib-SHAR Dataset This dataset is collected by scholars from the University of Milano-Bicocca, which is intended to identify a variety of "falling" activities. Data is collected from 30 participants ranging in age from 18 to 60 years old using an Android smartphone. During the data collection process, all participants must wear smart phones in their left and right pockets. The sensor signals is sampled at 50 Hz.

DSADS Dataset The DSADS dataset collects 19 activities performed for five minutes by 8 participants. We used nine of these activities. The entire signal length for each participant's activity is five minutes. The participants are asked to complete the activities in their style. The activities occur at three campus locations: the Bilkent University Sports Hall, the Electrical and Electronics Engineering Building, and a flat outdoor area. The sensor signals is sampled at 25 Hz.

MHEALTH Dataset The MHEALTH dataset contain recordings of body movements and vital signals from 10 participants with various characteristics. Each participants complete 12 exercises in an out-of-lab setting with no constraints. Three inertial measurement

units (IMUs) are attached to the chest, right wrist, and left ankle of the participants, respectively. In addition, the IMU on the chest provides two-lead ECG readings. The sensor signals is sampled at 50 Hz.

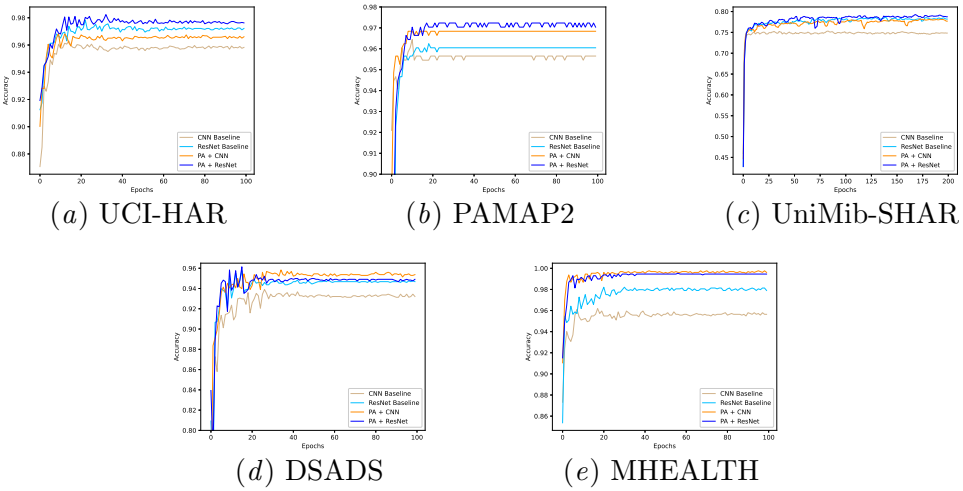
Table 3: Structure of 3-Layers Resnet and CNN (conv1-128 represents convolutional layer with 1 input channel and 128 out channels, $\longrightarrow$ represents input from the first layer to the second layer, FC represents full connection layers. The first line (bold type) is the parameters of Resnet, and the second line is the parameters of CNN).

Dataset	Structure
UCIHAR	**conv1-conv128$\longrightarrow$conv256-conv256$\longrightarrow$conv384-conv384$\longrightarrow$FC** conv64$\longrightarrow$conv128$\longrightarrow$conv256$\longrightarrow$FC
PAMAP2	**conv1-conv128$\longrightarrow$conv256-conv256$\longrightarrow$conv384-conv384$\longrightarrow$FC** conv128$\longrightarrow$conv256$\longrightarrow$conv384$\longrightarrow$FC
UniMib-SHAR	**conv1-conv128$\longrightarrow$conv256-conv256$\longrightarrow$conv384-conv384$\longrightarrow$FC** conv128$\longrightarrow$conv256$\longrightarrow$conv384$\longrightarrow$FC
DSADS	**conv1-conv128$\longrightarrow$conv256-conv256$\longrightarrow$conv384-conv384$\longrightarrow$FC** conv128$\longrightarrow$conv256$\longrightarrow$conv384$\longrightarrow$FC
MHEALTH	**conv1-conv128$\longrightarrow$conv256-conv256$\longrightarrow$conv384-conv384$\longrightarrow$FC** conv128$\longrightarrow$conv256$\longrightarrow$conv384$\longrightarrow$FC

4.2. Experimental Details

Platform All the models are trained/tested on two Nvidia K80 12GB GPU, Intel E5-2620 CPU, 64 GB memory.

Values of Hyperparameters Used in the Baseline Table 3 summarizes the values of the hyperparameters employed. The batch size in the UniMib-SHAR dataset is set to 128. In other datasets, the batch size is set to 64. In all datasets, the initial learning rate is 0.001. The default values are utilized for the other hyperparameters.



(*a*) UCI-HAR (*b*) PAMAP2 (*c*) UniMib-SHAR

(*d*) DSADS (*e*) MHEALTH

Figure 2: Test Accuracies with Different Models on Five Datasets.

Ablation Studies PA mechanism is critical. We execute a series of ablation tests to illustrate the performance of the proposed positional attenttion mechanism, the results are displayed in Table 4. To demonstrate the significance of positional information, we eliminate PA mechanism. As seen in Table 4, adding positional information improves the network greatly over the basic network. On the datasets UCIHAR, PAMAP2, UniMib-SHAR, DSADS, and MHEALTH the 3-layer Resnet is used to improve 0.62%, 0.46%, 1.12%, 0.64%, 0.18, and 1.44% respectively, and the 3-layer CNN is used to improve 0.75%, 1.21%, 3.00%, 2.11%, and 4.01% respectively. These experiments show that the PA mechanism is necessary for HAR classification.

Table 4: mAcc(%) of Models on Various Datasets.

Methods \ Dataset	UCIHAR	PAMAP2	UniMib-SHAR	DSADS	MHEALTH
Resnet + Positional Attention	**97.65**	**97.16**	**78.84**	94.85	99.45
Resnet Baseline	97.19	96.04	78.20	94.67	98.01
3-Layers CNN + Positional Attention	96.58	96.84	77.83	**95.40**	**99.64**
3-Layers CNN Baseline	95.83	95.63	74.83	93.29	95.63
Huang et al. (2021b)	96.40	95.67	77.55	94.44	98.76
Huang et al. (2021a)	94.68	94.86	75.42	94.52	96.68
Qian et al. (2019)	84.13	84.55	73.21	82.25	90.50
Teng et al. (2020)	95.23	94.59	76.19	94.29	95.51
Gao et al. (2020)	96.65	93.79	77.29	94.82	99.18

Comparison with Other Methods Our method is compared to both baseline and state-of-the-art methods. The results are shown in the Table 4 and Fig.2. Because feature-engineering-based machine learning approaches are difficult to scaled, we compare PA-HAR model with deep learning-based methods in this study. We follow five HAR classification methods as comparison, including Selective CNN (Huang et al. (2021b)), Shallow Convolutional (Huang et al. (2021a)), DDNN (Qian et al. (2019)), Local Loss CNN (Teng et al. (2020)) and DanHAR (Gao et al. (2020)). It can be seen from Table 4 that the PA method based on Resnet proposed by us has significantly improved on the five data sets compared with SOTA. Our method outperforms SOTA methods 1.0%, 1.49%, 1.29%, 0.58% and 0.46% on five datasets (UCIHAR, PAMAP2, UniMib-SHAR, DSADS, and MHEALTH) respectively.

4.3. Discussion

The accuracy of the PA-HAR model on the five datasets is improves respectively, as shown in Fig.2. Take PAMAP2 dataset as an example, as shown in Fig.2(b), our results show that the classification model with PA mechanism has a significant improvement in both convergence speed and accuracy. According to our experiment results, as seen in Table 4, we found that the accuracy of the model increased with the addition of positional information. The long-range dependency and sensor's cross-channel information are lost by convolution operation. It is reasonable that Eqn.(2) as well as Eqn.(3) extract the global temporal and sensor's channel information, and Eqn.(5) as well as Eqn.(6) use convolution to extract precisely positional information.

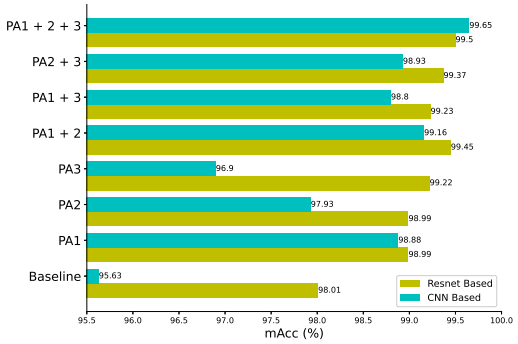


Figure 3: The Accuracy of PA Mechanism at Different Layers in MHEALTH Dataset (The PA1 represents insert the PA mechanism into after the first layer of network and PA1+2 represents insert the PA mechanism into after the first layer as well as second layer of network).

On the MHEALTH dataset, we perform ablation experiments to evaluate the effect of PA mechanism at various layers. As indicated in Fig.3, in CNN model, the PA mechanism should be added after the first, second, and third layers for maximum efficiency. It is due to the fact that the signal features at high-level have richer positional information. The convolution process maps single-channel data to multi-channel data, and PA extracts high-level information. In Resnet model, we add the PA mechanism to the first and second blocks. Although the highest accuracy is obtained by adding PA mechanism after each block, it causes an increase in parameters and the accuracy is only improved by 0.49%.

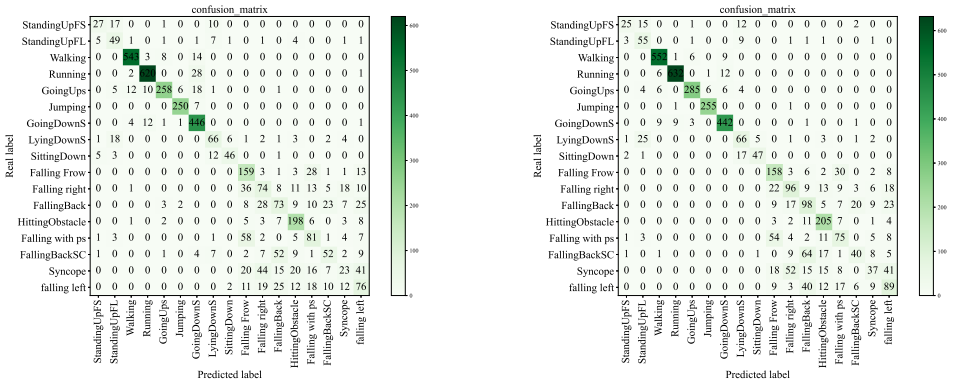
The experimental results show that our PA-HAR model is more robust than other methods. In LSTM-CNN related methods, for instance, DDNN (Qian et al. (2019)) is obviously inferior on UCIHAR and PAMAP2, and shallow convolution (Huang et al. (2021a)) is inferior on UniMib-SHAR. It is reasonable that LSTM has a forget gate, which will lose long-range dependency. And because the size of the CNN convolution kernel is constrained, cross convolution feature channel information cannot be taken into account overall. In the CBAM-attention related method (Gao et al. (2020)), the convolutional feature channel dimension in spatial attention mechanism of CBAM is squeezed to 1, leading to information loss. Second, CBAM encodes local spatial information using a convolutional process with 7×7 kernal size, which cannot cputure the global information. Furthermore, the CBAM-based HAR method is inferior on PAMAP2. In conclusion, our method perform well on multiple datasets.

Table 5: The Influence of H And W Attentions (The H is the temporal attention and W is the sensor's channel attention).

Network	Baseline	$+H$	$+W$	$+H+W$
mAcc	74.83%	76.69%	77.31%	**77.83%**

Table 6: The Execution Time Comparison on Different Methods.

Method	Resnet	Resnet + PA	Selective-CNN	Shallow-CNN	DDNN	Local-Loss	DanHAR
Time (s)	0.0481	0.0611	0.1364	0.9972	0.0468	**0.0452**	0.0637



(*a*) 3-layer CNN. (*b*) 3-layer CNN + PA.

Figure 4: Confusion Matrices on Unimib-SHAR Dataset.

As shown in Table 5, take UniMib-SHAR dataset as an example, we conclude that adding both long-range dependency and cross sensor's channel relationship can increase the classification accuracy of the network. when both attention mechanisms are added, we obtain the highest accuracy.

We conducted an execution time comparison for the test part of the MHEALTH dataset, which includes 1285 test data, and the results are shown in the following table. As shown in Table 6, though the proposed method is not the fastest, it does not have apparent disadvantages. Our method is not superior in time execution. The main reason is our PA mechanism, which increases the convolution and pooling processes, leading to an increase in execution time. Several results that perform better, such as the Local-Loss (Teng et al. (2020)) method, only use cosine similarity to make local losses upstream of the network. DDNN (Qian et al. (2019)) use fully connected mapping to high-dimensional space to extract the maximum mean difference.

According to our results, the propose model can be viewed as an additional step of operation based on CNN, which has better performance than CNN or other SOTA classification approaches. Referring to other HAR researches (Huang et al. (2021a); Gao et al. (2020)), to illustrate CNN's superiority in classification, we employ a confusion matrix to associate an explicit feature representation. The proposed model and the baseline CNN's confusion matrices on the Unimib-SHAR dataset for the HAR task are shown in Fig.4(a) and Fig.4(b). When comparing the PA-HAR approach to the baseline CNN for two similar activities, "fallingback" and "fallingright", it is clear that the PA-HAR method has fewer misclassifications.

5. CONCLUSION

The PA mechanism is used for the first time in a HAR situation in this study. Extensive experiments are conducted on five public HAR datasets by adding the attention mechanism in both the horizontal and vertical directions: the UCI-HAR dataset, the PAMAP2 dataset, the UniMibSHAR dataset, the DSADS dataset, and the MHEALTH dataset. For each dataset, we create a baseline Resnet and CNN. The experimental results have shown that the PA-HAR approach outperforms the state-of-the-art approaches by adding PA mechanism.

However, this method is not suitable for semi-supervised or unsupervised classification. Therefore, we want to expand the PA-HAR model for semi-supervised or unsupervised settings in the future, where the quantity of labeled training data is restricted.

Acknowledgments

This study is supported by the National Key Research & Development Program of China No. 2020YFC2007104, Natural Science Foundation of China (No. 61902379) and Youth Innovation Promotion Association CAS, Jinan S&T Bureau No. 2020GXRC030, the Funding for Introduced Innovative R&D Team Program of Jiangmen (Grant No.2018630100090019844) and Wuyi University Startup S&T research funding for senior talents (No. 504/5041700171).

References

K. Altun, B. Barshan, and O. Tunel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *Proceedings of the 4th international conference on Ambient Assisted Living and Home Care*, 2012.

O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, and C. Villalonga. mhealthdroid: A novel framework for agile development of mobile health applications. In *International Workshop on Ambient Assisted Living*, 2014.

I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. Attention augmented convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. Eye movement analysis for activity recognition using electrooculography. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):741–753, 2011. doi: 10.1109/TPAMI.2010.86.

Iveta Dirgová Luptáková, Martin Kubovčík, and Jiří Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, 22(5):1911, 2022.

Nidhi Dua, Shiva Nand Singh, and Vijay Bhaskar Semwal. Multi-input cnn-gru based human activity recognition using wearable sensors. *Computing*, 103(3):1–18, 2021.

Davide Figo, Pedro C. Diniz, Diogo R. Ferreira, and João M. P. Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Pers. Ubiquitous Comput.*, 14(7):645–662, 2010. doi: 10.1007/s00779-010-0293-9.

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

J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

W. Gao, L. Zhang, Q. Teng, H. Wu, and J. He. Danhar: Dual attention network for multimodal human activity recognition using wearable sensors. *arXiv e-prints*, 2020.

M. Guo, Z. Wang, N. Yang, Z. Li, and T. An. A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors. *IEEE Transactions on Human-Machine Systems*, 49(1):105–111, 2018.

N. Y. Hammerla, R. Kirkham, P. Andras, and T. Ploetz. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. page 65, 2013.

N. Y. Hammerla, S. Halloran, and T. Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv*, 2016.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

Wenbo Huang, Lei Zhang, Wenbin Gao, Fuhong Min, and Jun He. Shallow convolutional neural networks for human activity recognition using wearable sensors. *IEEE Trans. Instrum. Meas.*, 70:1–11, 2021a. doi: 10.1109/TIM.2021.3091990.

Wenbo Huang, Lei Zhang, Qi Teng, Chaoda Song, and Jun He. The convolutional neural networks training with channel-selectivity for human activity recognition based on sensors. *IEEE J. Biomed. Health Informatics*, 25(10):3834–3843, 2021b. doi: 10.1109/JBHI.2021.3092396.

Jongheon Jeong and Jinwoo Shin. Training CNNs with selective allocation of channels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3080–3090. PMLR, 09–15 Jun 2019.

H. Jie, S. Li, S. Gang, and S. Albanie. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017.

D. Linsley, S. Dan, S. Eberhardt, and T. Serre. Learning what and where to attend. 2018.

Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. Attnsense: multi-level attention mechanism for multimodal human activity recognition. In *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*, 2019.

Micucci, Daniela, Mobilio, Marco, Napoletano, and Paolo. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, 2017.

Hangwei Qian, Sinno Pan, and Chunyan Miao. Sensor-based activity recognition via learning from distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a.

Hangwei Qian, Sinno Pan, and Chunyan Miao. Sensor-based activity recognition via learning from distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. A novel distribution-embedded neural network for sensor-based activity recognition. In *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*, 2019.

A. Reiss and D. Stricker. Introducing a new benchmarked dataset for activity monitoring. In *International Symposium on Wearable Computers*, 2012.

Vijay Bhaskar Semwal, Neha Gaud, Praveen Lalwani, Vishwanath Bijalwan, and Abhay Kumar Alok. Pattern identification of different human joints for different human walking styles using inertial measurement unit (imu) sensor. *Artificial Intelligence Review*, (11), 2021a.

Vijay Bhaskar Semwal, Neha Gaud, Praveen Lalwani, Vishwanath Bijalwan, and Abhay Kumar Alok. Pattern identification of different human joints for different human walking styles using inertial measurement unit (imu) sensor. *Artificial Intelligence Review*, (11), 2021b.

A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, and Mads Mller Jensen. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Acm Conference on Embedded Networked Sensor Systems*, 2015.

Q. Teng, K Wang, L. Zhang, and J. He. The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. *IEEE Sensors Journal*, 20(13):7265–7274, 2020.

Kun Wang, Jun He, and Lei Zhang. Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. *IEEE Sensors Journal*, 19(17):7598–7604, 2019.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao. Rethinking and improving relative position encoding for vision transformer. 2021.

G. Yang, J. Liu, Y. Chen, X. Jiang, and H. Yu. Toselm: Timeliness online sequential extreme learning machine. *Neurocomputing*, 128(mar.27):119–127, 2014.

Ming Zeng, Haoxiang Gao, Tong Yu, Ole J. Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. Understanding and improving recurrent networks for human activity recognition by continuous attention. pages 56–63, 2018.