A Shallow Convolution Network Based Contextual Attention for Human Activity Recognition

Chenyang Xu[1,2], Zhihong Mao[1], Feiyi Fan[2], Tian Qiu[1], Jianfei Shen[2,3(✉)], and Yang Gu[2]

[1] Intelligent Manufacturing Department, Wu Yi Unitersity, Jiangmen, Guangdong 529020, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing100090, China
{fanfeiyi,shenjianfei,guyang}@ict.ac.cn
[3] Shandong Academy of Intelligent Computing Technology, Jinan 250102, Shandong, China

Abstract. Human activity recognition (HAR) is increasingly important in ubiquitous computing applications. Recently, attention mechanism are extensively used in sensor-based HAR tasks, which is capable of focusing the neural network on different parts of the time series data. Among attention-based methods, the self-attention mechanism performs well in the HAR field, which establish the correlation of key-query to fuse the local information with global information. But self-attention fails to model the local contextual information between the keys. In this paper, we propose a contextual attention (COA) based HAR method, which utilize the local contextual information between keys to guide learning the global weight matrix. In COA mechanism, we use $k \times k$ kernel to encode input signal to local contextual keys to extract more contextual information between keys. By fusing local key and query to generate global weight matrix, we can establish the correlation between local features and global features. The values are multiplied by the weight matrix to get a global contextual key, which include global contextual information. We combine the local key and global key to enhance feature's expression ability. Extensive experiments on five public HAR datasets, namely UCI-HAR, PAMAP2, UNIMIB-SHAR, DSADS, and MHEALTH show that the COA-based model is superior to the state-of-the-art methods.

Keywords: Contextual Attention (COA) · Deep Learning · Human Activity Recognition

1 Introduction

Human avtivity recognition (HAR) system can recognize various activities, such as running, walking, etc. HAR systems are used in numerous application scenarios, including medication intake, health monitoring and fitness

S. Longfei and P. Bodhi (Eds.): MobiQuitous 2022, LNICST 492, pp. 155–171, 2023.
https://doi.org/10.1007/978-3-031-34776-4_9

tracker. For example, in health monitoring field, HAR is utilized to help people analyze human behaviors (e.g. fall detection and Parkinson's disease assessment). Recently, utilizing wearable device to predict human activities becomes popular [1].

In the previous study, numerous traditional methods such as Logistic Regression, Decision Trees, Random Forest, Extreme Learning Machine (ELM), and Naive Bayesian approaches [2,3] are extensively used in HAR area, which achieve remarkable performance. Despite the traditional machine learning (ML) method in HAR providing various benefits, they need to extract features from the raw signal data manually and are usually complicated as well as time-consuming. While shallow feature is not good at classifying complex activities. As a result, the effectiveness of traditional ML approaches for classification tasks is heavily reliant on the efficacy of feature engineering.

The emerging deep learning related methods, such as Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Recurrent Neural Network (RNN), achieve enormous success in image segmentation, classification, object recognition, and natural language processing, etc. In HAR field, deep learning also achieves a great success, which overcomes the shortcoming of hand-crafted feature extraction approaches by using automated feature identification. 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 image using various machine learning methods.

With the popularity of self-attention mechanism, a HAR method based on self-attention mechanism is proposed [4], which extract useful information from sensor's signals by allocating their focus among signal features. The learned attention weights improve the ability to recognize target signals from background signals. However, self-attention mechanism mainly relies on the isolated pairwise query-key interaction to generate an attention matrix, as shown in Fig. 1(a), neglecting the rich contextual information between neighbour keys. For the signal feature of activity, as shown in Fig. 1(b), the change of the activity signal contains contextual information. Only considering the isolated key-query correlation can not extract enough features. Therefore, we need to consider the context features.

In this paper, we make use of the abundance of context features among input keys for a 2D sensor signal feature [5]. We propose a HAR method based on contextual attention (COA)(Figure 1(b)). The COA mechanism combines key context mining and self-attention learning over 2D signal feature into an unified architecture. In COA mechanism, we propose to build the COA in the following steps: (1) The $k \times k$ convolution is performed on all the neighbor keys to contextualize the local keys' representation. (2) The contextualized local keys and queries are concatenated by two successive 1×1 convolution operations to

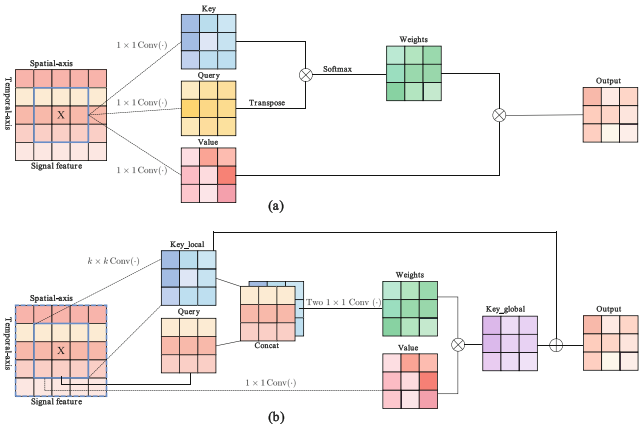


Fig. 1. Schematic comparison of (a) the traditional self-attention mechanism and (b) the proposed COA mechanism

generate global weight matrices. This process utilizes the reciprocal relationship between each query and all keys in self-attention to learn with the guidance of the local context information. (3) The learned attention weights matrix is employed to strengthen all the input values, forming global contextual information for the signal features. (4) We embed COA mechanism with CNN-HAR to enhance the ability of extracting contextual information. The contributions of our paper are summarized as follows.

- We propose a new COA mechanism for sensor-based HAR, which can make full use of input key's local contextual information to learn the global attention weights matrix. Then we multiply input values with attention weights matrix to obtain global keys.
- We employ two baseline networks (Resnet and 3-layers CNN), and add the COA mechanism to baseline network. We utilize large size kernel to capture local contextual information between keys. Then, we can employ local signal correlation to guide learning global feature correlation. The experiment result demonstrate that the local contextual features are significant.
- Extensive experiments on various benchmark datasets are carried out to demonstrate the higher performance of our proposed COA-HAR.

The following is a description of the paper's structure. We summarize HAR's related works in Sect. 2. The specifics of COA-HAR method are presented in Sect. 3. In Sect. 4, we detail experimental results obtained on five public HAR datasets, which are compared with the existing SOTAs. Moreover, several ablation studies about the COA-HAR method are provided. We describe conclusions in Sect. 5.

2 Related Works

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

2.1 Feature Engineering Approaches

In previous study, several state-of-the-art classification methods are employed for HAR. Awais et al. [6] use the SVM as a classifier to bulid a sensors-based activity classification system for elderly people, which can accurately recognize four primary activities of everyday life, including sitting, standing, walking, and lying. Ignatov et al. [7] employ k-nearest neighbors (KNN) approach as a classifier in a IMU-based HAR scene. Akhavian et al. [8] use five ML methods to classify workers' activities, and achieve a good performance. However, the ML methods need a huge amount of hand-crafted training data, and the absence of training data might restrict its effectiveness. These widely used ML methods usually suffer from significant limitations in HAR, such as progress slowly in the training and inference procedure, and poor generalization performance. With the development of extreme learning machine (ELM), a single hidden layer feed-forward neural network benefits from a fast running speed and a high degree of generalization [3]. The problem with ELM is that the algorithm parameters affect recognition accuracy and the randomly generated weights make it instable.

2.2 Deep Learning Approaches

Deep learning methods are able to extract features from raw signals and generate accurate predictions. Many deep learning-based algorithms have been applied to HAR field, including CNN, LSTM, CNN-LSTM, and attention-based CNN.

CNN is used to identify human activities. We process the raw sensor signals into 2D signal images and then send the signal image to the CNN model. Huang et al. [9] propose a channel selection method: the Expected Channel Damage Matrix (ECDM) is used to identify the contribution of each channel, and the channel with a low contribution rate is recovered in the training stage. In addition, the channel with a high contribution rate is reassigned to the position of the recovered low contribution channel. Cui et al. [10] employ multi-scale CNN to make an end-to-end classifier for univariate time series data. Andrey et al. [11] propose a CNN-based method for IMU-based HAR, where CNN is employed to extract local features, and the global signal features is derived by using statistical features. However, CNN is only limited to extracting local temporal correlation and sensor channel correlation information, and can not establish long-range dependency in temporal dimension and cross channel correlation in sensor channel dimension.

Another extensively used member of the DL family is RNN. The HAR is a time-series classification task. It is critical to capture temporal dependencies in the signal data. RNN is well-suited for this task. Some researchers employ RNN to classify human activities. For example, Chen et al. [12] propose

LSTM-based feature extractor for human activity categorization. Ullah et al. [13] propose another LSTM-based model. Firstly, the signal from the gyroscope and accelerometer are normalised. Secondly, the normalised signal is given to a stacked LSTM network to obtain an output, which would be placed into a softmax layer afterwards. Yu et al. [14] propose bi-directional LSTM-based model for HAR. Although the RNN can extract long sequence signals, the uncorrelated signals [15] in the sensors will affect the RNN.

A hybrid of CNN and RNN is used in recent HAR research. Qian et al. [14] integrates CNN-RNN models into an integrated framework to automatically extract long-range temporal features, statistical features, and cross sensor channel features, then merge them into an integrated feature map for HAR. Zeng et al. [15] propose temporal and sensor attention with CNN-LSTM models for human activity recognition, which adaptively focus on essential signals. Ma et al. [16] propose AttnSense for human activity recognition. AttnSense introduces the framework of merging attention mechanism with CNN and Gated Recurrent Units (GRU) to capture the interdependence of sensor signals in both the sensor channel and temporal dimension, which demonstrates benefits in prioritized sensor selection and enhances the comprehensibility. In CNN-RNN methods, although the long-range features can be extracted by RNN. However, uncorrelated signals affect the RNN, and then affect the effect of the model.

More recent research on HAR field uses a combination of CNN and attention mechanisms. Ramanujam et al. [17] employ ConvLSTM network with self-attention mechanism to extract temporal and sensor channel features from sensor signal. But self-attention mechanism only uses isolated query-key pairs to extract feature correlation, which ignores the context information around key. Moreover, in self-attention mechanism, only global correlation is considered and local contextual correlation is ignored.

3 Methdology

This section introduces the COA mechanism in detail. We denote multiple raw sensor signals to a predetermined window size as $S = \{s_1, s_2, \ldots, s_n\}$, where $S \in \mathbf{R}^{m \times n}$, S is signal image given to network, m is the length of the time series, and n denotes the sensor channel dimension. Based on self-attention, we propose the contextual attention (COA), a novel attention mechanism, which has three main operations to learn signal's feature.

- Local key generation: The COA mechanism uses $k \times k$ convolution to contextualize keys to capture contextual information between keys.
- Global key generation: The COA mechanism combines local keys and querys to form weights matrix. Then, we use weights matrix mutiple values to form global keys to strengthen global information.
- Local key and global key combination: After the above steps, the local key and global key are obtained. We combine the local key and global key as the final attention weights.

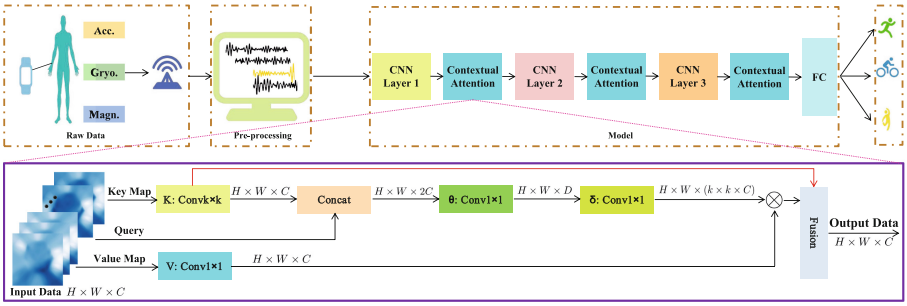


Fig. 2. Overview of Contextual Attention HAR Model Based on CNN

3.1 Rethinking Self-attention

In this part, let us rethinking the traditional self-attention [18]. The signal image is convolved to obtain a 2D feature map as $X^{C\times H\times W}$. Here C is the channel number, H is temporal axis of feature map, and W is sensor channel axis of feature map. We generate queries $Query = XW_{query}$, keys $Key = XW_{key}$, and values $Value = XW_{value}$ through X mutiple the embedding matrix $(W_{query}, W_{key}, W_{value})$ respectively. We can obtain each embedding matrix by 1×1 convolution. Then, the relation matrix $Mid \in \mathbb{R}^{C\times H\times W}$ between keys Key and queries $Query$ as is acquired as:

$$Mid = Key \otimes Query \tag{1}$$

where $\otimes$ is the operation of matrix multiplication, which calculates the pairwise correlation between each query. Mid represents the correlation information of each signal frame in the feature map.

After the above operation, we can obtain the attention matrix Att by normalizing the local matrix Mid with $Softmax$ activation function:

$$Att = Softmax(Mid) \tag{2}$$

The final output is obtained by multiplying the $Value$ with the Att weight matrix. The $Output$ can be written as:

$$Output = Value \otimes Att \tag{3}$$

These three transformations allow self-attention mechanism to aggregate local signal's temporal and sensor channel correlation, which helps the networks more accurately locate the objects of interest.

3.2 Contextual Attention Mechanism

The traditional self-attention mechanism establishes signal's temporal and sensor channel correlation. However, all correlation of query-key pairs in the traditional self-attention are learned on independent query-key pairs, which ignore the abundant contextual information between neighbor keys. The contextual information

is significant to signal feature extraction in sensor signal, which extract local temporal and sensor channel correlation in signal context. In order to solve this problem, we propose a novel attention mechanism named Contextual Attention (COA), which can combine contextual information exploit with self-attention mechanism in a unified framework.

Specifically, given a 2D feature map $X \in \mathbb{R}^{C\times H\times W}$ as the input. We define the queries and values as $Query = X$ and $Value = XW_{value}$, respectively. Firstly, COA mechanism utilizes group convolution with $k \times k$ kernel size across all the neighbor keys in $k \times k$ signal feature in order to spatially contextualize each key representation. The contextualized keys $Key_{\_local} \in \mathbb{R}^{C\times H\times W}$ represent the local neighbor keys' local contextual information, which capture local signal feature change. Secondly, we concatenate the $Key_{\_local}$ and $Query$, which is followed by two convolution operations with 1×1 kernel size. So, we can make the local signal feature and global signal feature fuse. Here θ is a convolution operation activated by a $ReLU$ function, and δ contains only a convolution operation.

$$Att = \delta(\theta([Key_{\_local}, Query])) \tag{4}$$

In each spatial position, the local attention matrix Att is generated by the $Query$ and the $Key_{\_local}$ instead of the independent query-key pairs. This method improves the information mining ability of self-attention mechanism. Finally, We obtain the global feature map $Key_{\_global}$ by aggregating values $Value$ with contextualized attention matrix Att.

$$Key_{\_global} = Value \otimes Att \tag{5}$$

As described above, $Key_{\_global}$ can capture the signal feature's global information. Therefore, we call the $Key_{\_global}$ as the global contextual information of signal feature. The result of our COA mechanism $Output$ can be written as:

$$Output = Key_{\_local} \oplus Key_{\_global} \tag{6}$$

where $\oplus$ represents add operation.

3.3 COA Based HAR Method

As this paper aims to investigate a attention mechanism to augment the convolutional features for HAR networks, we take 3-layers CNN architecture as baseline network. Then, we embed the COA mechanism with CNN-HAR network to demonstrate the method's effectiveness. To demonstrate the effectiveness of the COA mechanism in other networks, we also use a 3-layer Resnet as the baseline network and embed the COA mechanism after the first layer. Each layer in Resnet has two convolution processes. Each shortcut connection contains a convolution operation. We give the specific implementation steps of COA and embed the COA mechanism into the CNN baseline network, as shown in Fig. 2. In Fig. 3, we give the two baseline models and the model with the embedded COA mechanism.

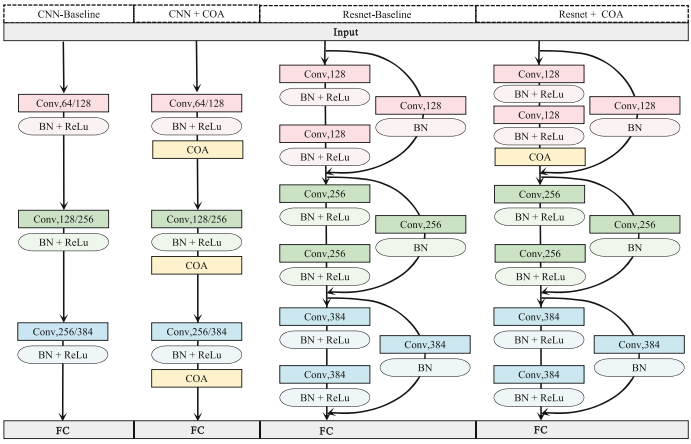


Fig. 3. Structure of 3-Layers Resnet and CNN (For UCIHAR dataset, the channel number of the first convolution layer is 64. For other datasets, the channel number is 128.)

4 Experiment

4.1 Datasets

In order to evaluate the effectiveness of the proposed method, we perform extensive experiments using five publicly available HAR datasets. The UCIHAR dataset [19], PAMAP2 dataset [20], UniMib-SHAR dataset [21], DSADS [22], and MHEALTH dataset [23] 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 datasets is shown in Table 2.

Table 1. Briefly Description of The Operation for Each 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: The UCIHAR dataset is collected using the embedded accelerometer and gyroscope of the Samsung Galaxy S II smartphone at a sampling frequency 50 Hz. The dataset is obtained from 30 volunteers aged 19 to 48

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

Dataset	The Class 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 with a Speed of 4 km/h (in Flat Position), Walking on a Treadmill with a Speed of 4 km/h (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

who wear a smartphone around their waist. The original signal data are preprocessed with a noise filter before being sliced according to the predetermined size of the sliding window. In this experiment, we utilize pre-processed data.

PAMAP2 Dataset: The PAMAP2 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. T100 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 9 Hz.

UniMib-SHAR Dataset: The UniMib-SHAR 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 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 25 Hz.

Table 3. The Key Map Value of COA Mechanism

Layers	Dataset				
	UCIHAR	PAMAP2	UniMib-SHAR	DSADS	MHEALTH
COA1 After The 1st CNN Layer	$K = 1$	$K = 3$	$K = 5$	$K = 3$	$K = 3$
COA2 After The 2nd CNN Layer	$K = 1$	$K = 3$	$K = 5$	$K = 5$	$K = 5$
COA3 After The 3rd CNN Layer	$K = 1$	$K = 3$	$K = 5$	$K = 5$	$K = 7$
COA1 After The 1st Resnet Layer	$K = 1$	$K = 5$	$K = 3$	$K = 5$	$K = 3$

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 50 Hz.

4.2 Experimental Platform

All the models are trained/tested on a single Nvidia RTX-3090 24GB GPU, Intel I5-10400 CPU, 32 GB memory. We use PyTorch deep learning library to implement all of the experiments (Table 4).

Table 4. Simple Description of The Neural Network Parameter

Dataset / Layers	UCI-HAR				
	conv	padding	stride	BN	ReLU
layer1	(6,1)	(1,1)	(3,1)	√	√
layer2	(6,1)	(1,1)	(3,1)	√	√
layer3	(6,1)	(1,1)	(3,1)	√	√
batch_size	64				
learning rate	0.001				

Dataset / Layers	PAMAP2				
	conv	padding	stride	BN	ReLU
layer1	(6,1)	(0,0)	(3,1)	√	√
layer2	(6,1)	(0,0)	(3,1)	√	√
layer3	(6,1)	(0,0)	(3,1)	√	√
batch_size	8				
learning rate	0.001				

Dataset / Layers	UniMib-SHAR				
	conv	padding	stride	BN	ReLU
layer1	(6,1)	(1,0)	(3,1)	√	√
layer2	(6,1)	(1,0)	(3,1)	√	√
layer3	(6,2)	(1,0)	(3,1)	√	√
batch_size	128				
learning rate	0.001				

Dataset / Layers	DSADS				
	conv	padding	stride	BN	ReLU
layer1	(3,3)	(2,0)	(1,1)	√	√
layer2	(3,3)	(2,0)	(1,1)	√	√
layer3	(3,3)	(2,0)	(1,1)	√	√
batch_size	64				
learning rate	0.001				

Dataset / Layers	MHEALTH				
	conv	padding	stride	BN	ReLU
layer1	(3,1)	(2,1)	(1,1)	√	√
layer2	(3,1)	(2,1)	(1,1)	√	√
layer3	(3,1)	(2,1)	(1,1)	√	√
batch_size	64				
learning rate	0.001				

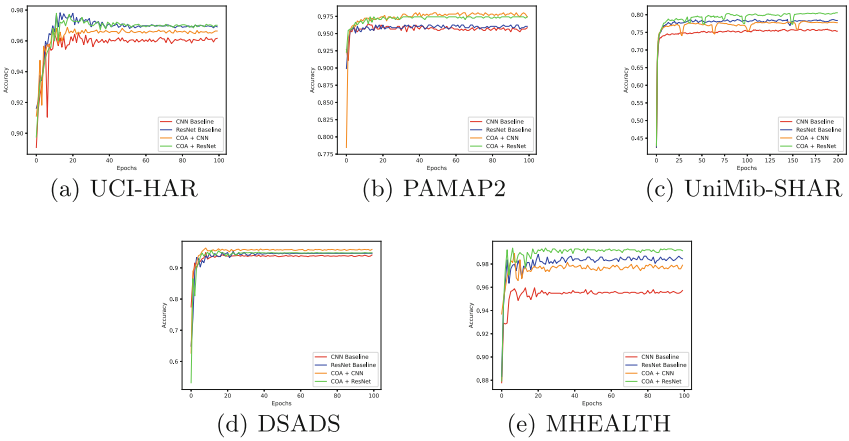
4.3 Values of Hyperparameters Used in the Baseline

The details of network structure are shown in the Fig. 3 summarizes the values of the hyperparameters employed. The K value in the COA mechanism in every dataset is shown in Table 3. The default values are utilized for the other hyperparameters.

4.4 Comparison with Other Methods

Our method is compared with both baseline and state-of-the-art methods. The results are shown in the Table 5 and Fig. 4. Because feature-engineering-based machine learning approaches are difficult to be scaled, we compare COA-HAR model with deep learning-based methods in this study. We follow five HAR classification methods as a comparison, including Selective CNN (Huang et al. [9]), Shallow Convolutional (Zhang et al. [24]), DDNN (Qian et al. [25]), Local Loss CNN (Teng et al. [26]) and DanHAR (Gao et al. [27]). It can be seen from Table 5 that the COA method based on 3-Layers CNN proposed by us improve significantly in the four data sets compared with SOTA. Note that the results with $*$ are directly cited from the references. Our method based on 3-layers CNN outperforms SOTA methods 1.16%, 0.23%, 0.50% on three datasets (PAMAP2, UniMib-SHAR, DSADS) respectively. And the COA method based on Resnet in UCIHAR, UniMib-SHAR, MHEALTH outperforms SOTA method 0.34%, 2.75%, 0.01%.

- Selective CNN [9]: a state-of-the-art Resnet-based model with 3 convolutional blocks. This method uses selective convolution to select the contribution of each channel through the ECDM matrix, replacing low-contribution channels with high-contribution channels. We reproduce the model by following the architecture described in the paper.
- Shallow Convolutional [24]: a state-of-the-art CNN-based model with 3 convolutional blocks and GCN block. This method uses GCN to capture the information between channels so that each channel is interconnected. We reproduce the model by following the architecture described in the paper.
- DDNN model [25]: a state-of-the-art LSTM and CNN model. This method uses LSTM to capture sensor channel and temporal features and CNN to capture the temporal and sensor channel connection. We reproduce the model by following the architecture described in the paper.
- Local Loss CNN [26]: a state-of-the-art CNN-based model with 3 convolutional blocks and local loss block. This method uses local loss block to optimize loss in the upstream part. We reproduce the model by following the architecture described in the paper.
- DanHAR model [27]: a state-of-the-art Resnet-based model with 3 convolutional blocks and temporal and sensor channel attention. This method uses Convolutional Block Attention Module (CBAM) [28] attention based Resnet for HAR. We reproduce the model by following the architecture described in the paper.



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

(d) DSADS (e) MHEALTH

Fig. 4. Test Accuracy with Different Models on Five Datasets

4.5 Ablation Studies

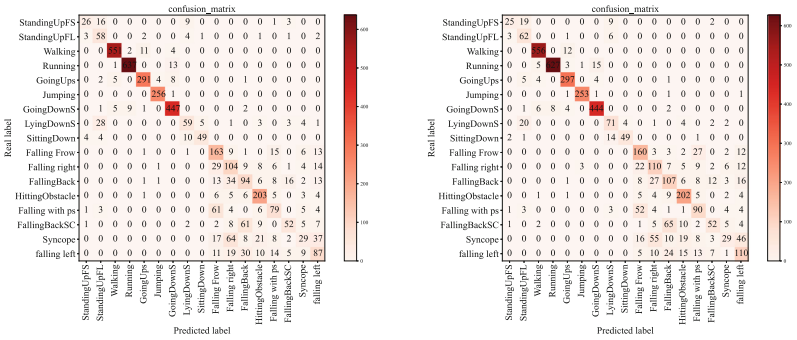
In order to demonstrate the COA mechanism is critical, we execute a series of ablation experiments. As seen in Table 5, adding contextual information improves greatly the basic network. On the datasets UCIHAR, PAMAP2, UniMib-SHAR, DSADS and MHEALTH datasets, the 3-layer CNN with COA mechanism improve 0.53%, 0.73%, 2.18%, 1.96%, and 2.18% respectively. And in Resnet with COA mechanism, the accuracy improves 0.04%, 0.67%, 2.04%, 0.12% and 0.77% on UCIHAR, PAMAP2, UniMib-SHAR, DSADS, and MHEALTH respectively. These experiments show that the COA mechanism is necessary for HAR classification.

Table 5. mAcc(%) of Models on Various Datasets

Methods	Dataset				
	UCIHAR	PAMAP2	UniMib-SHAR	DSADS	MHEALTH
Resnet + COA	96.99	97.36	**80.35**	94.78	**99.19**
Resnet Baseline	96.95	96.00	78.31	94.66	98.42
CNN + COA	96.58	**97.80**	77.78	**95.81**	97.75
CNN Baseline	96.05	95.62	75.60	93.85	95.57
CNN + Self-Attention	96.20	93.56	76.71	95.31	96.73
Huang et al. [9]	96.40	95.67	77.55	94.44	98.76
Zhang et al. [24]	94.68	94.86	75.42	94.52	96.68
Qian et al. [25]	84.13	84.55	73.21	82.25	90.50
Teng et al. [26]	95.23	94.59	76.19	94.29	95.51
Gao et al. [27]	96.65	93.79	77.29	94.82	99.18
Xu et al.* [29]	–	–	80.02	–	–
Huang et al.* [30]	**97.35**	92.14	78.65	–	–

4.6 Visualizations

The COA-HAR approach may be seen as an additional stage of operation based on CNN that beats CNN and other SOTA classification methods in accuracy. The COA mechanism improves the accuracy of activity classification by extracting the local context signal features. Table 5 shows that the COA method based on 3-Layers CNN is superior to other SOTA methods in four datasets. As other HAR studies [24] [27], we use confusion matrix to illustrate the advantages of CNN in classification. The proposed model and the baseline CNN's confusion matrices on the UniMib-SHAR dataset for the HAR task are shown in Fig. 5(a) and Fig. 5(b). When comparing the COA-HAR method to the baseline CNN for two similar activities, "GoingUps" and "Walking", it is clear that the COA-HAR method has fewer misclassifications. For falls in different directions, take "Falling right" and "Falling back" as example. Although the activities are very similar, the classification accuracy is still improve.



(a) Resnet Baseline.

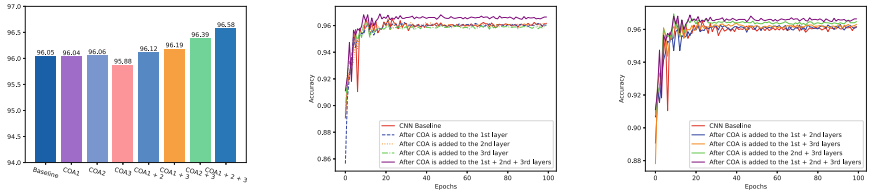
(b) Resnet + COA.

Fig. 5. Confusion Matrices on UniMib-SHAR Dataset

4.7 Discussion

Where is the Best Position to Insert COA Mechanism? On the UCI-HAR dataset, we perform ablation experiments to evaluate the effect of COA block at various layers. As indicated in Fig. 6, the COA block should be added after the first, second, and third layers for maximum efficiency. It is due to the fact that high-level contextual information is encoded after the convolution process. The COA mechanism combined with multi-layer feature information can get better feature expression capability. As a result, adding COA mechanism at every convolution layer allows the network to acquire more meaningful information for activity recognition.

Can COA Mechanism Reduce the Depth of Network? According to our results in Table 5, especially on UCIHAR dataset, the 3-layers Resnet performance is similar to 3-layers CNN. As shown in Fig. 4(c), the 3-layer CNN with



(a) (b) (c)

Fig. 6. Accuracy of COA Mechanism at Different Layers

COA mechanism is able to achieve a comparable or even better accuracy than the 3-layers Resnet. As a result, COA-HAR model may achieve similar classification precision than multi-layers model (Fig. 7).

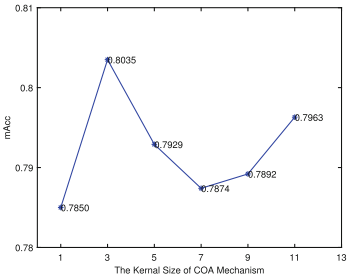


Fig. 7. Effect of The K in COA mechanism

How the Value of K ***Influence the Accuracy?*** According to our results on UniMib-SHAR dataset in Fig. 6, the result in Resnet is shown when the $K = 3$ the accuracy is the best. According to the dataset's characteristics, the activity signals are periodic or transient, such as "Running" or "Falling". Therefore, we infer that although the large K value can extract a larger range of contextual information, it will introduce irrelevant signal features (signals outside the period or after the activity occurs).

Does the COA-HAR Model is Robust? We find that the proposed COA-HAR method exhibits better robustness when compare with the other methods. In LSTM-CNN related methods, for instance, DDNN [25] is obviously inferior in UCIHAR and UniMib-SHAR, and Layer-wise CNN [26] is inferior in UniMib-SHAR. One reason could be that DDNN is a unified framework where the signal of different parts are learned together. The signal feature learned by LSTM might encode some noise in the data itself (such as irrelevant signal components) [15]. And because the size of the CNN convolution kernel is constrained, cross-channel information cannot be taken into account. In the Triple Cross-Domain related method [31], the accuracy is inferior in PAMAP2 dataset. The reason maybe that

the sensor dimension, temporal dimension, and sensor channel dimension are captured by three attention branches, and they ignore the correlation between the dimensions. In conclusion, the COA-HAR model improved significantly compared with the baseline in multiple datasets. It performs well in four datasets, proving that our method is robust.

What is the Difference Between the COA Mechanism and Self-attention? As can be seen in Fig. 1, the COA mechanism fully considers the contextual information of signal features, which is significant for activity recognition. For periodic activity such as "Cycling", the HAR method based on self-attention only consider the isolated key-query correlation information. Therefore, the correlation between the signal at a certain time and the global signal can be considered, and the local contextual information within the periodic activity will be ignored. Unlike image data, sensor data is temporally correlated, and the sensor's channels are also correlated. The COA mechanism can correlate local context information and global information, which is very important for activity classification. It can be seen from the results in the Table 5 that the COA-HAR method is significantly improved compare with the HAR method based on self-attention, which proves the importance of contextual information in self-attention.

5 Conclusion

In this paper, we concentrate on exploring contextual information between the local signal feature and global signal feature for HAR problem with a CNN model. The COA-HAR method is proposed, which exploits the neighbour key's contextual information to guide self-attention to capture more correlation information of signal feature context. The COA mechanism first captures the local context among neighbor keys, then the COA mechanism is employed to guide self-attention to exploit the global context information. This method combines context exploiting and self-attention mechanism into an uniform framework, improving the capacity of signal feature recognition. Our experimental results show that the COA mechanism is critical for increasing the performance of CNN architectures.

Acknowledgments. This study is supported by the National Key Research & Development Program of China No. 2020YFC2007104, Natural Science Foundation of China (No.61902377), 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), the Wuyi University Startup S&T research funding for senior talents 2019 (No. 504/5041700171).

References

1. Wang, Z., Jiang, M., Yaohua, H., Li, H.: An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors. IEEE Trans. Inf Technol. Biomed. **16**(4), 691–699 (2012)
2. Woodford, B.J., Ghandour, A.: An information retrieval-based approach to activity recognition in smart homes. In: Hakim, H., et al. (eds.) ICSOC 2020. Lecture Notes in Computer Science, vol. 12632, pp. 583–595. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-76352-7_51
3. Donghui, W., Wang, Z., Chen, Y., Zhao, H.: Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. Neurocomputing **190**, 35–49 (2016)
4. Betancourt, C., Chen, W.H., Kuan, C.W.: Self-attention networks for human activity recognition using wearable devices. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1194–1199. IEEE (2020)
5. Li, Y., Yao, T., Pan, Y., Mei, T.: Contextual transformer networks for visual recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2022
6. Awais, M., Chiari, L., Ihlen, E.A.F., Helbostad, J.L., Palmerini, L.: Physical activity classification for elderly people in free-living conditions. IEEE J. Biomed. Health Inf. **23**(1), 197–207 (2018)
7. Ignatov, A.D., Strijov, V.V.: Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. Multimedia Tools Appl. **75**(12), 7257–7270 (2016)
8. Akhavian, R., Behzadan, A.H.: Smartphone-based construction workers' activity recognition and classification. Autom. Constr. **71**, 198–209 (2016)
9. Huang, W., Zhang, L., Teng, Q., Song, C., He, J.: The convolutional neural networks training with channel-selectivity for human activity recognition based on sensors. IEEE J. Biomed. Health Inf. **25**(10), 3834–3843 (2021)
10. Cui, Z., Chen, W., Chen, Y.: Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:1603.06995 (2016)
11. Ignatov, A.: Real-time human activity recognition from accelerometer data using convolutional neural networks. Appl. Soft Comput. **62**, 915–922 (2018)
12. Chen, Y., Zhong, K., Zhang, J., Sun, Q., Zhao, X.: LSTM networks for mobile human activity recognition. In: 2016 International Conference on Artificial Intelligence: Technologies and Applications, pp. 50–53. Atlantis Press (2016)
13. Ullah, M., Ullah, H., Khan, S.D., Cheikh, F.A.: Stacked lstm network for human activity recognition using smartphone data. In: 2019 8th European Workshop on Visual Information Processing (EUVIP), pp. 175–180. IEEE (2019)
14. Yu, S., Qin, L.: Human activity recognition with smartphone inertial sensors using bidir-lstm networks. In: 2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp. 219–224. IEEE (2018)
15. Zeng, M., et al.: Understanding and improving recurrent networks for human activity recognition by continuous attention. In: Proceedings of the 2018 ACM International Symposium on Wearable Computers, pp. 56–63 (2018)
16. Ma, H., Li, W., Zhang, X., Gao, S., Lu, S.: Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In: IJCAI, pp. 3109–3115 (2019)
17. Ramanujam, E., Perumal, T., Padmavathi, S.: Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. IEEE Sens. J. **21**(12), 13029–13040 (2021)

18. Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10076–10085 (2020)
19. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Bravo, J., Hervas, R., Rodriguez, M. (eds.) IWAAL 2012. Lecture Notes in Computer Science, vol. 7657, pp. 216–223. Springer, Cham (2012)
20. Reiss, A., Stricker, D.: Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th International Symposium on Wearable Computers, pp. 108–109. IEEE (2012)
21. Micucci, D., Mobilio, M., Napoletano, P.: Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. Appl. Sci. **7**(10), 1101 (2017)
22. Altun, K., Barshan, B., Tunçel, O.: Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recognit. **43**(10), 3605–3620 (2010)
23. Banos, O., et al.: mHealthDroid: a novel framework for agile development of mobile health applications. In: Pecchia, L., Chen, L.L., Nugent, C., Bravo, J. (eds.) IWAAL 2014. LNCS, vol. 8868, pp. 91–98. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13105-4_14
24. Huang, W., Zhang, L., Gao, W., Min, F., He, J.: Shallow convolutional neural networks for human activity recognition using wearable sensors. IEEE Trans. Instrum. Meas. **70**, 1–11 (2021)
25. Qian, H., Pan, S.J., Da, B., Miao, C.: A novel distribution-embedded neural network for sensor-based activity recognition. In: IJCAI, vol. 2019, pp. 5614–5620 (2019)
26. Teng, Q., Wang, K., Zhang, L., He, J.: The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. IEEE Sens. J. **20**(13), 7265–7274 (2020)
27. Gao, W., Zhang, L., Teng, Q., He, J., Hao, W.: Danhar: dual attention network for multimodal human activity recognition using wearable sensors. Appl. Soft Comput. **111**, 107728 (2021)
28. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
29. Shige, X., Zhang, L., Huang, W., Hao, W., Song, A.: Deformable convolutional networks for multimodal human activity recognition using wearable sensors. IEEE Trans. Instrum. Meas. **71**, 1–14 (2022)
30. Huang, W., Zhang, L., Wu, H., Min, F., Song, A.: Channel-equalization-HAR: a light-weight convolutional neural network for wearable sensor based human activity recognition. In: IEEE Transactions on Mobile Computing (2022)
31. Tang, Y., Zhang, L., Teng, Q., Min, F., Song, A.: Triple cross-domain attention on human activity recognition using wearable sensors. In: IEEE Transactions on Emerging Topics in Computational Intelligence (2022)