Expert Systems With Applications 198 (2022) 116764

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Human activity recognition using wearable sensors by heterogeneous convolutional neural networks

Chaolei Han [a], Lei Zhang [a,*], Yin Tang [a], Wenbo Huang [a], Fuhong Min [a], Jun He [b]

[a] School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, 210023, China
[b] School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, China

ARTICLE INFO

Keywords:
Sensor
Grouped convolution
Activity recognition
Deep learning
Heterogeneous convolution

ABSTRACT

Recent researches on sensor based human activity recognition (HAR) are mostly devoted to designing various network architectures to enhance their feature representation capacity for raw sensor data. In this paper, we focus on strengthening the vanilla convolution without adjusting the model architectures in HAR scenario. Inspired by the idea of grouped convolution, we propose a novel heterogeneous convolution for activity recognition task, where all filters within a specific convolutional layer are separated into two uneven groups. Specifically, the sensor input is down-sampled into a low-dimensional embedding, which is then convolved by one filter group to recalibrate normal filters within the other group. The two filter groups can complement each other, which is very beneficial for augmenting the receptive field of sensor signals for HAR task. Extensive experiments are conducted on several benchmark HAR datasets, which consists of OPPORTUNITY, PAMAP2, UCI-HAR, USC-HAD as well as the Weakly Labeled HAR dataset. The results show that the baseline models can be significantly improved. Our heterogeneous convolution is simple and can easily be integrated into standard convolutional layers without increasing extra parameters and computational overhead. Finally, the actual operation of heterogeneous convolution is evaluated on an embedded Raspberry Pi platform.

1. Introduction

With the rapid technical development of Internet of Things and sensing technology, various motion sensors can be embedded into smart devices such as phones and watches to record people's motion information. Due to obvious advantages, e.g., lower cost, smaller size, and flexible deployment, smart sensing devices embedded with inertial measurement units (IMUs) such as accelerometers and gyroscopes provide a better alternative to unobtrusively perform activity recognition task (Nweke et al., 2018; Ronao & Cho, 2016; Wang, Cang et al., 2019). The recognizing systems can monitor and analyze people's behaviors with sensory time series to improve the quality of their daily life. During the past decade, sensor based human activity recognition (HAR) (Dang et al., 2020; Wang, He et al., 2021) has gained a lot of attention due to its rapid growth in a large variety of application domains such as interactive games, smart homes, and health care. For example, HAR can offer a smart medical assistance by identifying the action being undertaken by a patient. Monitoring activities of daily living (ADLs) of patients with chronic diseases such as obesity and cardiovascular has played a vital role in smart healthcare (Ogbuabor & La, 2018; Wang, Cang et al., 2019). According to the report from World Health Organization (WHO), the major cause of obesity can be attributed to the lack of physical exercise. Automatic activity recognition systems are able to assist the physicians to effectively monitor and analyze daily living habits of those patients, hence offering proper diagnosis and treatment. In smart home scenario (Feng et al., 2017), activity recognition also can be applied in many surveillance tasks such as fall detection in the elderly. Due to the serious problem of aging population, physical inactivity of the elderly people will not only affect their living quality, but also bring financial burden to the societies and individuals. Recent years have witnessed the success of assistive HAR systems using sensors, Internet of Healthcare Things (Zhou et al., 2020), and machine learning techniques, which can build long-term and elderly-friendly environment. Activity recognition enables people to have a real-time interaction with game devices (Lara & Labrador, 2012), which leads to an immersive entertainment experience. Due to the enrichment of sensing data, HAR has become an active research topic in ubiquitous computing scenario, which may provide reliable support for the development of various human-centric services or applications.

* Correspondence to: Nanjing Normal University, No. 2 Xuezhe Road, Qixia District, Nanjing, Jiangsu Province, China.
E-mail addresses: chaoleihan98@gmail.com (C. Han), leizhang@njnu.edu.cn (L. Zhang), yinntag@gmail.com (Y. Tang), WenboHuang1002@outlook.com (W. Huang), minfuhong@njnu.edu.cn (F. Min), jhe@nuist.edu.cn (J. He).

https://doi.org/10.1016/j.eswa.2022.116764
Received 20 July 2021; Received in revised form 17 January 2022; Accepted 25 February 2022
Available online 11 March 2022

Various machine learning algorithms such as Naive Bayes, K-nearest neighbors, and support vector machine (Bulling et al., 2014) have been extensively explored in HAR. These traditional machine learning techniques for sensor-based HAR mainly focused on the design of shallow hand-crafted features, which are domain-specific or task-dependent. Recent breakthroughs in deep learning technique that can automatically extract relevant representative features have significantly pushed the latest state-of-the-art in HAR. For example, Yang et al. (2015) proposed a deep convolutional neural network (CNN) to handle HAR problem with multichannel time series, where the non-handcrafted features learned by the CNN are task-dependent. Hammerla et al. (2016) detailed various deep, convolutional, and recurrent network architectures across three benchmark HAR datasets that contain motion data recorded by wearable sensors. The current research focus in HAR is undergoing a transition from feature engineering to network engineering. That is to say, how to adjust the network architectures to be optimal for generating better feature representations? Accordingly, more research efforts are devoted to hand-designed network architectures such as CNNs (Albawi et al., 2017; Huang et al., 2021; Kalchbrenner et al., 2014), residual networks (ResNets) (He et al., 2016) or their diverse variants, which inevitably requires too much human labor.

Actually, it is very difficult for one practitioner to determine what are the most optimal model architectures for their HAR applications. In the paper, in order to avoid tuning complex network architectures, we propose a novel heterogeneous convolutional network, which aims to strengthen the basic convolution to produce discriminative features in HAR scenario. Similar to grouped convolution (Chollet, 2017; Howard et al., 2017; Krizhevsky et al., 2012; Zhang et al., 2018), the core idea behind the method is to divide the filters of a specific convolutional layer into two groups but unevenly. The filters within each group are applied to a sensor input in a heterogeneous way. To be specific, the sensor input is first transformed into a low-dimensional embedding via down-sampling operation. The low-dimensional embeddings are processed by one filter group, which are then utilized to recalibrate normal filters within the other group. The heterogeneous convolutions can complement each other, which is very beneficial to augment the receptive field of sensor signals for HAR application. Let us briefly comment on a number of closely related background details, which strongly motivate our research in this paper. Generally speaking, standard CNNs contain three kinds of layers: convolution layer, pooling layer, and fully connected layer, which may be stacked hierarchically to form a deep classifier for activity recognition. Actually, the convolution operations with fixed-size kernel can be directly implemented along temporal dimension of sensor signals to extract discriminative activity features. Fig. 1 illustrates the waveforms of acceleration signals from the 'walking' and 'running' activities. Considering different activity speeds at 'walking'/'running', they will be more discriminative at different time scales. Conventional convolution usually has a fixed kernel size, which can only detect the signal fluctuations at a fixed time scale. To fill this gap, Lee et al. (2017) ensembled multiple CNN architectures that have different kernel sizes to extract features at multiple time scales. Because this multi-kernel CNN architecture will require expensive computation, it is very impractical for real-time or lightweight HAR by deep models on wearable and mobile devices. Furthermore, when a larger time scale is desirable, a pooling operation will be inserted between two convolutional layers, which inevitably causes information loss because of the subsampling process from time series. To handle this issue, Xi et al. (2018) adopted a dilated convolution method to time series, which utilizes dilated kernels rather than standard kernels to expand the receptive field (i.e., time length) without information loss. The dilated convolution nearly requires no extra computation, because it only injects empty elements into the standard kernel. But the time scale that an individual dilated convolution may explore is inadequate as well. The varying time scale is a critical concern in ubiquitous HAR scenario. To fill the gap, different from previous works, we propose a heterogeneous two-stream CNN architecture, which may handle different time scales for activity recognition. The main contributions of the proposed method are three-fold:

1. In order to avoid the shortcoming of fixed time scale in normal convolution, we for the first time propose a new heterogeneous two-stream CNN architecture to encode contextual information of sensor time series from different receptive field sizes, which can generate more discriminative activity features at different time scales for activity recognition.
2. The proposed two-stream convolution is a Plug-and-Play block, which can be easily integrated into the existing deep models for HAR without increasing any extra memory and computation cost or changing other network hyperparameters.
3. We conduct extensive experiments on several public HAR datasets consisting of OPPORTUNITY, PAMAP2, UCI-HAR, USC-HAD as well as the Weakly Labeled HAR dataset to evaluate how varying such heterogeneous convolutions affect the overall recognition performance. Ablation studies verify that our heterogeneous two-stream convolutions with different subsampling rates are able to better extract activity features at different time scales. The actual performance is evaluated via running the HAR systems on an embedded platform.

The rest of this paper is organized as follows. In Section 2, recent related deep learning and HAR researches are reviewed. The overview of the proposed method is presented in Section 3. We perform extensive experiments on various public HAR datasets in Section 4. Discussion is conducted in Section 5. Finally, a conclusion is made.

2. Related work

During recent years, deep learning (LeCun et al., 2015) has made remarkable advances in the field of HAR. At the earliest time, Zeng et al. (2014) presented a CNN-based feature extraction approach to capture the scale invariance and local dependency characteristics of sensor time series for HAR task, which outperforms traditional hand-crafted feature approaches. In order to handle HAR using 1D time-series signal, Ronao and Cho (2016) proposed a deep CNN, which can automatically extract the inherent temporal local dependency, scale invariance, and hierarchical features of activities. Ignatov (2018) exploited CNNs for local feature extraction together with simple hand-designed statistical features, which can preserve contextual information about the global form of time series. In multimodal HAR scenario, Chen and Xue (2015) proposed a sophisticated CNN, which contains three convolutional layers with 18, 36, 24 filters followed by 2×1 max-pooling layers each respectively. In order to extract the association between two adjoining pairs of sensor axes, they adopted a 12×2 filter at the first layer. Jiang and Yin (2015) proposed a novel CNN with two layers, in which they adopt filters of 5×5 followed by 4×4 and 2×2 average-pooling layers, respectively. In particular, raw sensor signals are transformed into a two-dimensional activity image for classification. Ma et al. (2019) proposed a new deep model called AttnSense for multimodal HAR tasks, which combines attention module with a CNN and a Gated Recurrent Units (GRU) network to highlight more important sensor modalities or time intervals. The attention-based deep model can improve the interpretability of deep model behaviors. Using self-supervised idea, Haresamudram et al. (2021) presented a new Contrastive Predictive Coding (CPC) approach for HAR, which is able to preserve high recognition performance when there is only a very small number of labeled activity samples. Because it is very laborious for human to annotate activity data from a long sensor sequence, this method demonstrates its practical use due to the scarcity of annotation data. To protect user-sensitive information, Xiao et al. (2021) proposed a deep learning-based federated learning system, which is able to maintain satisfactory recognition accuracy and meanwhile preventing privacy leakage. Luo et al. (2021) introduced a binarized convolutional network for realtime HAR, in which the dilated convolution is used to enlarge receptive field and improve its potential capturing capability for time series. This work will effectively reduce latency in resource-constrained

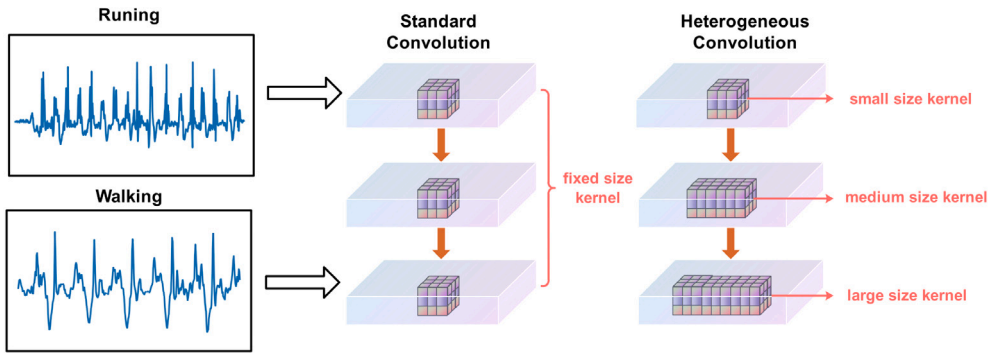


Fig. 1. The comparison between normal convolution and heterogeneous convolution scaling for different activities.

mobile devices, which may better support computation-intensive deep models in ubiquitous HAR scenario. Although deep learning has a great potential to automatically extract effective features from raw sensor signals, there is still no clear consensus on what is the optimal network architecture across a large variety of HAR tasks. Various hand-crafted network designs in HAR field require too much human labors.

During the past decade, deep network architecture design has made considerable advances in the field of computer vision. As an earliest work, Krizhevsky et al. (2012) presented AlexNet, which produces significant performance improvement by sequentially stacking a specific number of convolutional layers. Simonyan and Zisserman (2014) further introduced VGGNet, which stacks more layers with smaller filters compared with AlexNet. In order to tackle the gradient vanishing problem, He et al. (2016) proposed ResNet, where an identity map is used as skip connection to generate extremely deep networks. Szegedy et al. (2015) designed GoogLeNet by using carefully designed Inception modules, which contains multiple parallel paths of sets of specialized filters to extract features. However, it is hard to train such deep neural networks in a resource-constrained platform. Thus, the idea of grouped convolution has been firstly proposed to divide the number of channels in half, which can drastically reduce the number of computations to obtain output feature maps. Liu et al. (2020) proposed an efficient self-calibrated convolution, which performs all the convolutions over the input in an uneven way. However, the idea of grouped convolution is rarely to be seen in HAR scenario. Different from all above-mentioned HAR researches that focus on adjusting hand-designed network architectures, we first propose to use the idea of grouped convolution to design powerful feature extractor for HAR, which can augment the basic convolutional module to generate more rich feature representations.

3. Model

In this section, we will explore the idea of grouped convolution, and present the proposed framework of heterogeneous convolution for our activity recognition challenge. Different from imagery data (He et al., 2016), raw sensor signals (Zeng et al., 2018) need to be first preprocessed, which refers to noise removal (Rudin et al., 1992), signal segmentation, and resampling processes. In particular, it is a crucial step to segment sensor time series for the subsequent activity recognition procedures. Because the sliding window approach can be easily implemented and require no preprocessing, it is ideally suitable for real-time HAR applications, where the sensor time series can be divided into continuous fixed-length samples with an overlap rate. The heterogeneous sensor values are then normalized into zero mean and unit variance by subtracting the mean and dividing by the standard variance.

Without loss of generality, a normal convolutional layer is composed of a set of filters $K = [k_1, k_2, \dots, k_{\hat{C}}]$, in which k_i denotes the ith filter. By standard convolution, the sensor input $X = [x_1, x_2, \dots, x_C]$ can be transformed into an output $Y = [y_1, y_2, \dots, y_{\hat{C}}]$, where $\hat{C}$ denote the number of filters. For notational convenience, omitting the filter size and bias term, we can formulate the output feature map at channel i as:

$$y_i = k_i * X = \sum_{j=1}^{C} k_i^j * x_j \tag{1}$$

in which '*' denotes convolutional operation. As a result, each output feature map can be computed by the summation across all channels. Repeating the Eq. (1) $\hat{C}$ times, we are able to obtain the final output Y (Kalchbrenner et al., 2014; Kim, 2017; Liu et al., 2020). Actually, the receptive field within a specific convolution layer is mainly predetermined by the fixed kernel size. The small filter is computationally efficient, but at the same time it is hard to capture long-range contextual information, which may lead to less discriminative feature maps. In order to avoid the above shortcoming, based on the idea of grouped convolution, we use heterogeneous convolution to strengthen vanilla convolution for improving the performance of HAR.

For grouped convolution (Chollet, 2017; Howard et al., 2017; Krizhevsky et al., 2012), all filters are divided into parallel branches, which are performed in a homogeneous way. The outputs from each branch are then concatenated to produce the final output. Based on the idea of grouped convolution, without loss of generality, the proposed approach divides all convolutional filters into two parts, yet in a heterogeneous way, each part is unevenly treated which is in charge of specific functionality. The heterogeneous convolution can generate different receptive fields, which leads to a better understanding of global contextual information. Fig. 2 shows the flowchart to recognize human activities. For the sake of simplicity, we assume that there is no change in the number of channels during the whole process, *i.e.* , $C = \hat{C}$. Given a group of filter sets K with shape (C, C, T, S) where T and S are height and width of the filters respectively, we first divide it into four parts, e.g., K_1, K_2, K_3, K_4 as illustrated in Fig. 2. Here the channel number C is divided by 2. Different from grouped convolution, each part with shape $(\frac{C}{2}, \frac{C}{2}, T, S)$ has its specific functionality. Based on above four parts of filters, the input X is divided into two parts X_1, X_2, each of which with shape $(\frac{C}{2}, T, S)$ is then sent into a special branch, which can aggregate contextual information at different scales. To be specific, there are two different scale spaces: one is an original scale space that shares the same resolution with original sensor input; the other is a down-sampling small-scale space, which can be easily obtained by pooling (Bruckstein et al., 2003; Sun et al., 2017) operation. We detail how to perform the heterogeneous convolution as follows:

For the former, because we only perform down-sampling on sensor time series, both the pooling size and stride in small-scale space are $(r, 1)$, which can be formulated as:

$$P_1 = Pool_r(X_1) \tag{2}$$

Based on K_1, the feature transformation is performed on P_1 :

$$X_1' = Up(F_1(P_1)) = Up(P_1 * K_1) \tag{3}$$

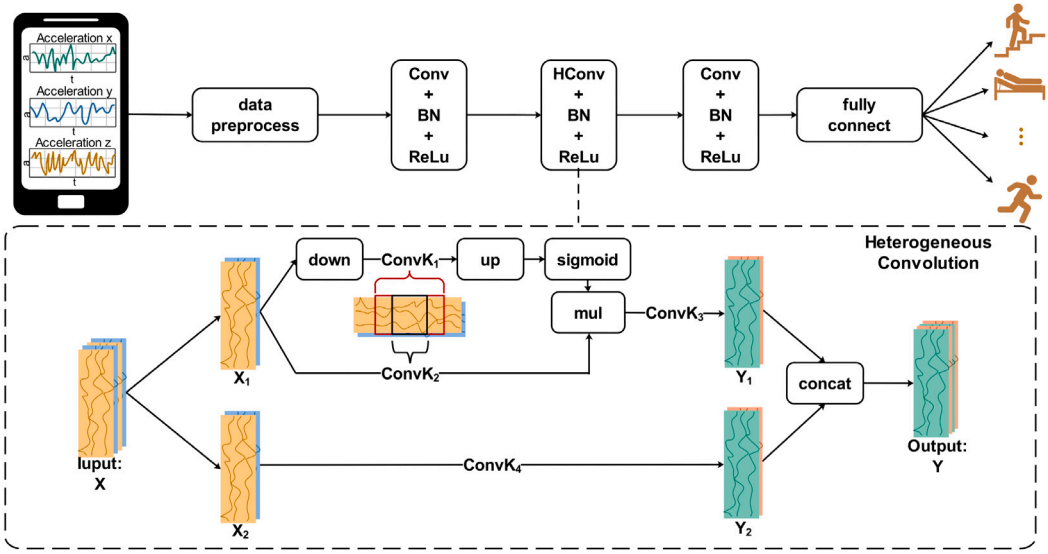


Fig. 2. The overview of heterogeneous convolution.

where the bilinear interpolation (Gribbon & Bailey, 2004; Kirkland, 2010; Mastyło, 2013) operator Up(·) is utilized to resize the feature maps of small-scale space to the original resolution. The operation in the original scale still can be formulated as follows:

$$X_2' = F_2(X_1) = X_1 * K_2 \tag{4}$$

Because of the pooling operation, the receptive field of K_1 is r times larger than of K_2's, where r is a hype-parameter of pooling operation. Therefore, the small-scale space can guide the feature transformation process in the original one (Hu, Shen et al., 2018; Huang et al., 2017; Liu et al., 2020; Woo et al., 2018). We use X_1' residuals to recalibrate the weights for guiding, which could be beneficial:

$$Y_1' = X_2' \cdot \sigma(X1') \tag{5}$$

where σ denotes sigmoid activation (Sibi et al., 2013) and '·' is element-wise operator. The output Y_1 in this branch can be written as :

$$Y_1 = F_3(Y_1') = Y_1' * K_3 \tag{6}$$

All in all, the process in the first branch is presented in Eqs. (2) to (6). For the latter, this process in the second branch can be given as:

$$Y_2 = F_4(X_2) = X_2 * K_4 \tag{7}$$

which is normal convolution to preserve original contextual information of sensor time series. Finally, we concatenate both the intermediate outputs Y_1, Y_2 to obtain final desired output Y.

4. Experiments

We divide this section into three parts. In part one, we detail the five HAR datasets used, which consist of OPPORTUNITY (Chavarriaga et al., 2013), PAMAP2 (Reiss & Stricker, 2012), UCI-HAR (Anguita et al., 2012), USC-HAD (Zhang & Sawchuk, 2012) and our Weakly Labeled HAR dataset (Wang et al., 2019b). In part two, the data preprocessing and network architectures of our baselines are presented. In part three, the effectiveness and efficiency of the proposed method are evaluated on several benchmark HAR datasets. All the models are trained by an Adam optimizer to minimize the cross-entropy loss function. Our algorithm is run by using the deep learning framework PyTorch with CPU Intel i7 6850k, 64 GB memory, and an NVIDIA 3090 GPU with 24 GB video memory.

4.1. Dataset description

4.1.1. OPPORTUNITY dataset (Chavarriaga et al., 2013)

Daniel et al. from the University of Sussex built this HAR dataset in a sensor-rich environment, which consists of 15 wireless and wired networked sensor systems. The sensor system has 72 sensors of 10 modalities within it. In a breakfast scenario, 17 kinds of activities were recorded from four subjects. On body, each subject was equipped with wearable sensor nodes for inferring human activities. The sampling frequency was set to 30 Hz.

4.1.2. PAMAP2 dataset (Reiss & Stricker, 2012)

This PAMAP (Physical Activity Monitoring for Aging People) dataset was built by researchers from the Department of Augmented Vision German Research Center of Artificial Intelligence. Nine subjects with 8 males and 1 female took part in the data collection, whose ages range from 27 to 30 years old. All subjects wore 3 Inertial Measurement Units (IMUs) and a heart-rate monitor, which were attached to the dominant's arm, ankle, and chest respectively. This dataset contains 18 kinds of activities, which consist of 'walking', 'cycling', 'rope jumping', etc. The sampling rate was set to 100 Hz. The PAMAP2 dataset is publicly available.

4.1.3. UCI-HAR dataset (Anguita et al., 2012)

This dataset was built by the researchers from the University of California Irvine for evaluating various machine learning algorithms on HAR task. The thirty subjects with their ages between 19 and 48 years took part in the data collection. All subjects were equipped with Samsung Galaxy S2 placed on their waists. Under a supervised scenario, each subject performed the six types of activities of daily living, which consists of 'walking', 'standing', 'lying', 'walking upstairs', and 'walking downstairs'. The sensor signals were recorded at a frequency of 50 Hz by triaxial angular velocity and acceleration sensors.

4.1.4. USC-HAD dataset (Zhang & Sawchuk, 2012)

This dataset was designed as a benchmark for comparing various algorithms particularly in healthcare scenario, which contains 12 kinds of activities such as 'walking forward', 'walking left', 'walking right', 'sleeping', 'sitting', performed by 7 males and 7 females. The 14 participants ranged in age from 21 to 49 and height between 160 cm and 185 cm. They recorded sensor signals by a sensing platform called MotionNode, where a triaxial accelerometer, a triaxial magnetometer, and a triaxial gyroscope are incorporated. All the participants wore the MotionNode on their front right hip for 5 trials. It took an average of 6 h for each subject to complete the entire dataset.

Fig. 3. Data collection process. These images from left to right are 'going upstairs', 'going downstairs', 'jogging', 'jumping' and 'walking'.

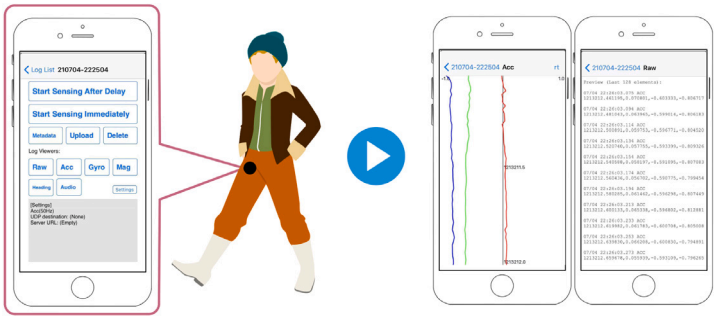


Fig. 4. The user interface of HascLogger.

Table 1
Data statistics of Weakly Labeled HAR dataset.

Activity	Label	Number
Go upstairs	0	4270
Go downstairs	1	9605
Jumping	2	3234
Jogging	3	4632
Total	–	21 741

4.1.5. Weakly labeled HAR dataset (Wang et al., 2019b)

This Weakly Labeled HAR dataset was collected by 10 volunteers, who used a triaxial accelerometer embedded in iPhone 7 placed in their right trouser pocket. Fig. 3 illustrates each volunteer's data collection process. Data collection was done through an application software called HascLogger (Kawaguchi et al., 2011), which is able to collect motion data by iPhone. The user interface of the application is shown in Fig. 4. Using this software, we can setup the following configurations such as measurement data (i.e., acceleration), sampling frequency, and the time of measurement. Activity data can be collected in real-time at a sampling rate of 50HZ. In a supervised scenario, each volunteer performs five types of daily activities consists of 'walking', 'jumping', 'jogging', 'going upstairs' and 'going downstairs', where the activity walking is seen as the background activity, while the other four activities are the recognized target activities. Each specific activity was repeated four times. A fixed-length window of 2048 is slid over sensor readings, which corresponds to 40.96 s. As a result, it produces overall 21,741 activity samples. The statistics of different types of activity samples are shown in Table 1. Due to inexact segmentation, each weakly labeled 2048-length activity sample may contain one or multiple target activities, as well as background activities.

4.2. Data preprocessing and network architecture description

The details of data preprocessing such as the sampling rate, window size, and overlapping rate are illustrated in Table 2. In the experiments, each dataset is divided into three parts, in which the training set, validation set, and test set account for 70%, 10%, and 20% of the total samples respectively. The other hyper-parameters such as training epochs, batch size, and learning rate are also listed in Table 2.

Both baseline backbone networks are used to evaluate the effectiveness of the heterogeneous convolution. One contains three convolutional layers and a fully connected layer, where batch normalization and the non-linear activation ReLU follow each convolution and max-pooling performs down-sampling before the final fully connected layer. The heterogeneous convolution is used to replace standard convolution within each intermediate convolutional layer as shown in Table 3. The other is residual network with three residual blocks, each of which includes two convolution layers, and a fully connected layer is inserted at the end to output classification. The network architecture of the ResNet is shown in Table 4.

4.3. Heterogeneous convolution VS vanilla convolution

4.3.1. Performance on OPPORTUNITY dataset

Fig. 5 shows the performance improvements of both baselines caused by the heterogeneous convolution. The classification performance, the number of parameters and FLOPs are presented for quantitative comparison. As can be seen from Table 5, the baseline CNN achieves an F1 score of 90.19%, while the heterogeneous convolution surpasses it by 0.81% with smaller FLOPs and almost the same number of parameters. Similarly, the heterogeneous ResNet could provide an improvement of F1 score by 0.45% over the corresponding baseline. Moreover, we compare the proposed method and the recent state-of-the-art approaches on OPPORTUNITY dataset. Table 5 shows that our method significantly surpasses (Kim, 2020)'s interpretable CNN by 4.15% in terms of the F1 score. The proposed heterogeneous convolution is superior to Hammerla et al. (2016) and Hu, Chen et al. (2018)'s results by 2.15% and 2.4% respectively. The performance of our heterogeneous convolution is very close to Ordóñez and Roggen (2016)'s 91.7% which uses DeepConvLSTM, but our method only uses CNN alone that requires no LSTM module.

4.3.2. Performance on PAMAP2 dataset

Fig. 6 compares the classification performance of both baselines with/without heterogeneous convolution. The evaluations are also performed according to FLOPs and the number of parameters. It is clear that the heterogeneous convolution consistently boosts classification performance under two different settings. Especially, for CNN and

Table 2
Summary of setup for datasets.

Dataset	Setting						
	Numbers of activities	Frequencies of sampling (Hz)	Sizes of windows	Overlap	Epoch	Batch size	Learning rate
OPPORTUNITY	18	30	30	50%	200	512	1e−4
PAMAP2	12	100	171	78%	200	512	5e−4
UCI-HAR	6	50	128	50%	200	256	5e−4
USC-HAD	12	100	512	50%	200	256	1e−4
Weakly Labeled HAR	4	50	2048	50%	200	256	3e−4

Table 3
Architecture of CNN with heterogeneous convolution(HC).

Structrue	Dataset				
	OPPORTUNITY	PAMAP2	UCI-HAR	USC-HAD	Weakly Labeled HAR
Layer1	3 × 3,S = 2,64	6 × 1,S = (3,1),64	6 × 1,S = (3,1),64	6 × 1,S = (3,1),64	6 × 1,S = (3,1),128
Layer2	3 × 3,S = 1,64	3 × 1,S = 1,64	3 × 1,S = 1,64	5 × 1,S = 1,64	3 × 1,S = 1,128
Layer3	3 × 3,S = 2,128	6 × 1,S = (3,1),128	6 × 1,S = (3,1),128	6 × 1,S = (3,1),128	6 × 1,S = (3,1),256
MaxPool	–	–	–	6 × 1	6 × 1
Fully connected layer					

Table 4
Architecture of ResNet with heterogeneous convolution(HC).

Structrue	Dataset				
	OPPORTUNITY	PAMPA2	UCI-HAR	USC-HAD	Weakly labeled HAR
Block1	$\begin{bmatrix} 3\times3, 64 \\ \text{HC, } 64 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 64 \\ \text{HC, } 64 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 64 \\ \text{HC, } 64 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 64 \\ \text{HC, } 64 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 64 \\ \text{HC, } 64 \end{bmatrix}$
Block2	$\begin{bmatrix} 3\times3, 128 \\ \text{HC, } 128 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 128 \\ \text{HC, } 128 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 128 \\ \text{HC, } 128 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 128 \\ \text{HC, } 128 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 128 \\ \text{HC, } 128 \end{bmatrix}$
Block3	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 256 \\ 3\times1, 256 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 256 \\ 3\times1, 256 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 256 \\ 3\times1, 256 \end{bmatrix}$	$\begin{bmatrix} 6\times1, 256 \\ \text{HC,256} \end{bmatrix}$
MaxPool	4 × 3	–	–	6 × 1	6 × 1
Fully connected layer					

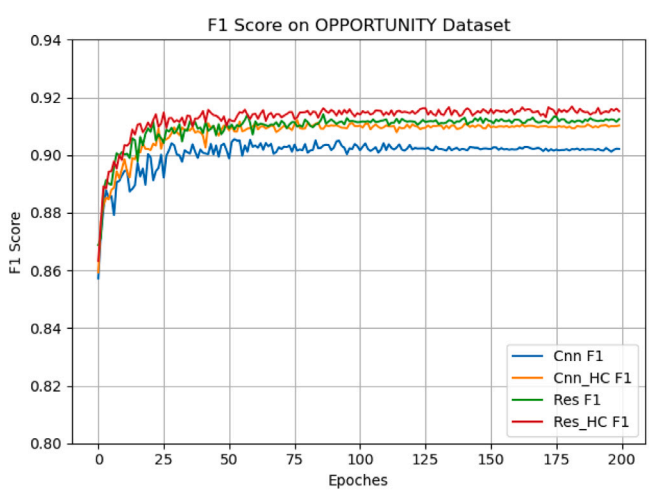


Fig. 5. F1 score on **OPPORTUNITY** dataset.

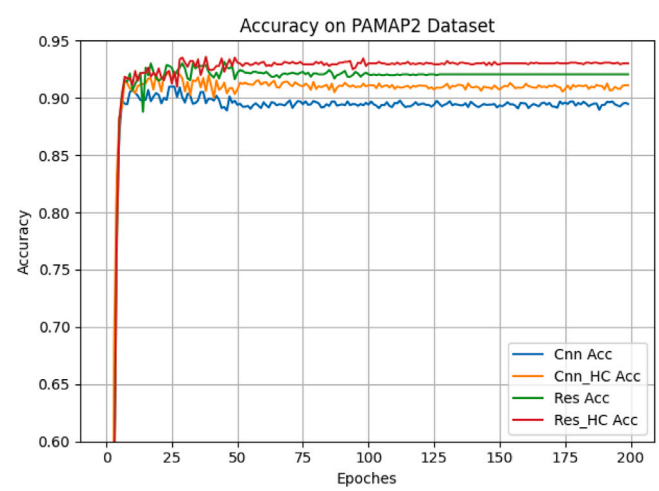


Fig. 6. Accuracy on **PAMAP2** dataset.

residual network, the models with heterogeneous convolution outperform the counterparts by a significant margin, which produce an accuracy improvement of 1.55% and 1.37% respectively. As shown in Table 5, the proposed method has almost the same parameters and fewer FLOPs when compared with both baselines. Compared with recent state-of-the-art methods on PAMAP2 dataset, the heterogeneous convolution method significantly surpasses (Ma et al., 2019) and Zeng et al. (2018)'s methods using attention mechanism by 3.67% and 3.01% respectively. It is significantly superior to Chen et al. (2019)'s method that uses a multi-agent spatial–temporal attention model by 2.64% as well. Especially, our method performs better than recent (Xia et al., 2021)'s method using a multiple-level domain adaptive learning model by 0.92%.

4.3.3. Performance on UCI-HAR dataset

We use the experiment settings listed in Table 2 to train our models. Results are shown in Fig. 7. It is clear that the proposed heterogeneous models outperform both baselines by 0.54% and 0.85% respectively, especially under almost the same complexity levels. Table 5 also compares our model with other state-of-the-art networks. It can be seen that the heterogeneous two-stream CNN performs better consistently, which surpasses (Ronao & Cho, 2016)'s method using standard convolution

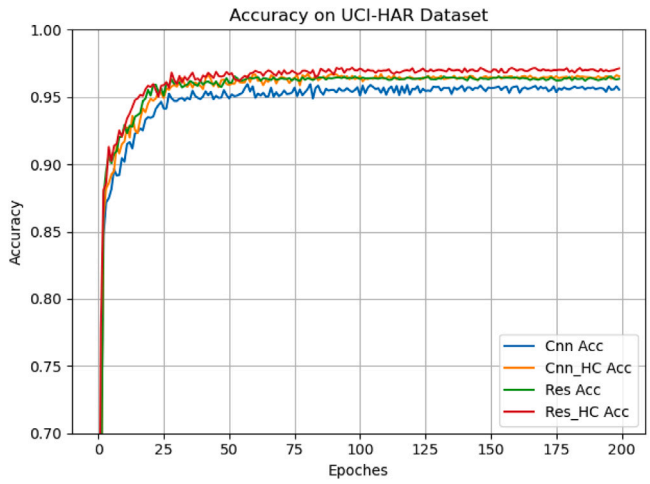


Fig. 7. Accuracy on **UCI-HAR** dataset.

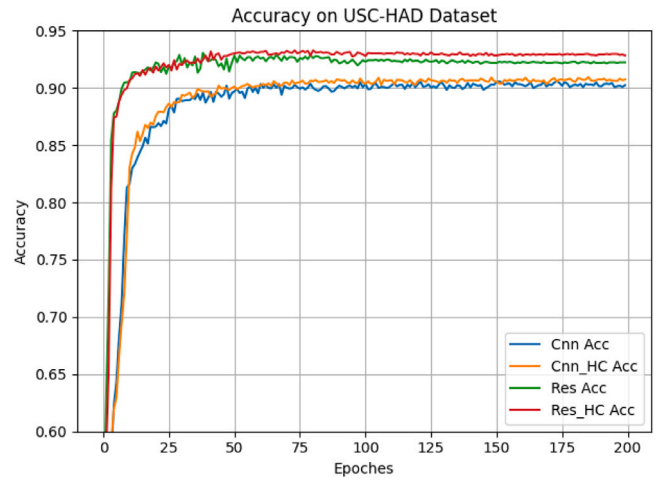


Fig. 8. Accuracy on **USC-HAD** dataset.

by 1.26%. Our method is significantly superior to Dong et al. (2021)'s method using hesitant fuzzy belief framework and Khan and Ahmad (2021)'s method using multi-head self-attention mechanism by 1.6% and 1.61% respectively. The heterogeneous CNN also provides a 0.64% performance gain over (Ignatov, 2018)'s result, which uses CNN for activity feature extraction together with handcrafted statistical features. We believe this is because the heterogeneous convolution can capture contextual information at different time scales, which is very beneficial for recognizing human activities from sensory data.

4.3.4. Performance on USC-HAD dataset

The detailed classification results for recognition on USC-HAD dataset are illustrated in Table 5. As one can see, our model obtains an accuracy of 90.67% and 93.49% in both cases, outperforming the baselines by 0.3% and 1.25%. Performance analysis shows that the heterogeneous model can lead to considerable improvements without increasing computational overhead. Recently, Kwon et al. (2018) have shown that temporal structure can be injected to distribution-based features of sensory data, which is able to effectively improve recognition performance in standard sliding window-based HAR chain. The heterogeneous two-stream CNN is obviously superior to their method by 7.01%. Moreover, our method provides a 1.79% performance gain compared with Bi et al. (2020)'s method that uses dynamic active learning. It also leads to an accuracy improvement of 2.42% over (Li et al., 2021)'s method that using federated representation learning framework. Comparing with Haresamudram et al. (2020)'s method which uses mask reconstruction based self-supervision, the heterogeneous convolution shows an accuracy improvement of 2.24% (see Fig. 8).

4.3.5. Performance on Weakly Labeled HAR dataset

Fig. 9 shows that the heterogeneous convolution can outperform both baselines in the weakly supervised task, which is able to consistently produce higher classification accuracies. Results from Table 5, it can be clearly observed that the accuracies of two baselines are 90.51% and 92.28% respectively, while our method surpasses them by 0.86% and 1.52% with smaller FLOPs. Benefitting from larger receptive field, the heterogeneous two-stream convolution shows a better feature extraction capability in weakly supervised activity recognition, which significantly surpasses (Wang et al., 2019b)'s method using soft attention mechanism and Gao et al. (2021)'s method using selective kernel convolution by 3.76% and 0.95% respectively. The experiment results verify that our method can not only be suitable for traditional supervised learning tasks, but also perform well on Weakly Labeled HAR dataset.

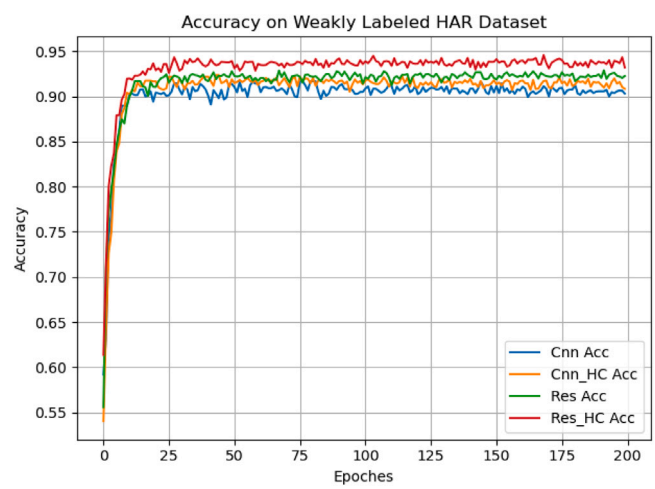


Fig. 9. Accuracy on **Weakly** Labeled HAR dataset.

5. Discussion

In this section, we perform extensive ablation experiments to analyze the effect of the heterogeneous convolution by changing the down-sampling pooling and its hyper-parameter r. In order to better understand what the advantage of the proposed model is, we visually show the influence of down-sampling operation on the Weakly Labeled HAR dataset. Moreover, we provide visualized analysis to the channel weights on PAMAP2 dataset, which can be utilized to evaluate the influence of different sensor modalities placed on different locations of human body. Finally, we evaluate the actual-time operation of the proposed model in an embedded system for efficient consideration.

5.1. The optimal down-sampling rate

As indicated in Section 3, we introduce the down-sampling operation to realize heterogeneous convolution within one branch, which has been proven very effective for improving vanilla convolution. In this part, we continue to explore how the down-sampling rate r in the heterogeneous convolution influences the recognition performance. Fig. 10 shows the classification performance with various down-sampling rates used in the heterogeneous convolution. For each dataset, we perform 10 runs to calculate their mean value and standard deviation. As can be seen from Fig. 10, on OPPORTUNITY dataset, $r = 3$ has almost the

Table 5
Accuracy&Parameters&FLOPs of models on various datasets.

Model	Dataset				
	OPPORTUNITY	PAMAP2	UCI-HAR	USC-HAD	Weakly labeled HAR
CNN	90.19%*&1.05M&99M	89.44%&1.17M&66M	95.65%&0.15M&11M	90.37%&0.15M&39M	90.51%&0.36M&239M
CNN+HC	91%*&1.05M&88M	90.99%&1.17M&61M	96.19%&0.15M&10M	90.67%&0.15M&35M	91.37%&0.36M&219M
ResNet	91.1%*&1.55M&327M	91.6%&1.37M&257M	96.16%&0.41M&44M	92.24%&0.42M&128M	92.28%&0.79M&263M
ResNet+HC	91.55%*&1.55M&306M	92.97%&1.37M&245M	97.01%&0.42M&42M	93.49%&0.42M&123M	93.8%&0.79M&243M
Others' Results	89.4%* (Hammerla et al., 2016)	89.30% (Ma et al., 2019)	95.41% (Dong et al., 2021)	91.25% (Haresamudram et al., 2020)	90.04% (Wang et al., 2019b)
	91.7%* (Ordóñez & Roggen, 2016)	89.96% (Zeng et al., 2018)	96.37% (Ignatov, 2018)	91.7% (Bi et al., 2020)	92.85% (Gao et al., 2021)
	89.15%* (Hu, Chen et al., 2018)	90.33% (Chen et al., 2019)	95.75% (Ronao & Cho, 2016)	86.48% (Kwon et al., 2018)	–
	87.4%* (Kim, 2020)	92.05% (Xia et al., 2021)	95.4% (Khan & Ahmad, 2021)	91.07% (Li et al., 2021)	–

1 Number with '*' means F1 score.

Table 6
Performances comparison between Maxpool or Avgpool.

Dataset	Maxpool	Avgpool	Performance
OPPORTUNITY	✓	×	91.39%*
	×	✓	91.55%*
PAMAP2	✓	×	92.22%
	×	✓	92.97%
UCI-HAR	✓	×	96.32%
	×	✓	97.10%
USC-HAD	✓	×	93.01%
	×	✓	93.49%
Weakly Labeled HAR	✓	×	92.54%
	×	✓	93.80%

1 Number with '*' means F1 score.

same F1 score with $r = 4$ and $r = 5$ but has smaller standard deviation. In the case of UCI-HAR, USC-HAD and Weakly Labeled HAR dataset, the optimal down-sampling rate r is 4. For PAMAP2 dataset, $r = 5$ can attain its peak value. The optimal down-sampling rate also depends on sliding window size. In order to prevent the case that the receptive field is out of scope for sliding window, we do not use larger down-sampling rates.

5.2. Maxpool vs. avgpool

In addition to the above down-sampling rate, we continue to explore the effect of different types of pooling operation on classification performance. Without loss of generality, keeping the other hyper-parameters unchanged, we shift all the max-pooling operations in heterogeneous convolution to the average-pooling operations to show potential performance difference. Table 6 shows that the average-pooling operation consistently outperforms max-pooling operation by 0.16%, 0.75%, 0.78%, 0.48%, 1.26% on OPPORTUNITY, PAMAP2, UCI-HAR, USC-HAD and Weakly Labeled HAR dataset respectively. For sensor signals, it can be attributed to the fact that average-pooling can better represent the overall strength of a feature by passing gradients through all indices, while gradient flows through only the max index in max-pooling.

For the Weakly Labeled HAR dataset, Fig. 11 visually shows the original receptive field and the various receptive fields caused by different down-sampling rates within the heterogeneous convolution. It can easily be observed that the yellow box that crops the original receptive field cannot locate the interesting target activity very well. Under the down-sampling rate $r = 5$, the receptive field represented by the purple box contains too much background noise, which could deteriorate final classification performance. In the case of $r = 4$, the receptive field with red box is more appropriate to match the target activity, which is well in line with our above results. All in all, it can be clearly observed that the receptive fields produced by heterogeneous convolution can more precisely locate the target activities and do not expand to the background areas too much, which is very helpful for discovering more integral target activities due to their flexible sizes.

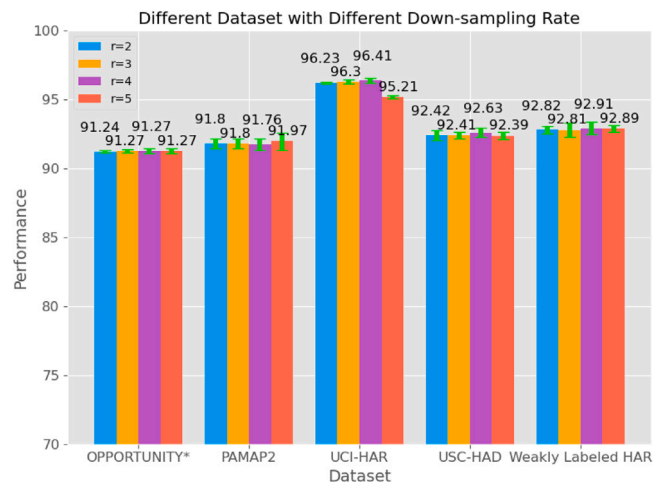


Fig. 10. Performance of different down-sampling rates for different datasets. F1 Score used on OPPORTUNITY dataset and accuracy used on others.

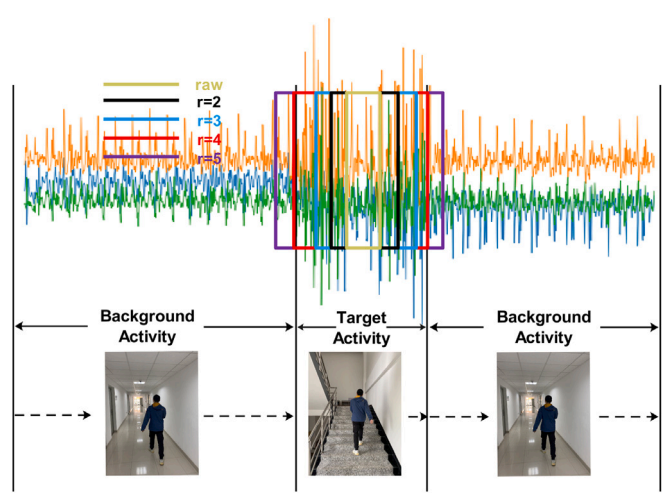


Fig. 11. Visualization of Weakly Labeled HAR dataset. The yellow box represents the receptive field of filter in original scale space while the other boxes represent the receptive fields of filters in small latent spaces which ratio are 2, 3, 4 and 5 respectively.

In order to show why the heterogeneous convolution is helpful for improving vanilla convolution, we provide visualization analysis to the channel weights in multimodal HAR scenario. We perform the ablation experiment on PAMAP2 dataset, where three IMUs are placed on different body parts of one subject consisting of chest, arm, and ankle respectively. One heart rate monitor is also used. We choose accelerometer, gyroscope and magnetometer as input. Fig. 12 shows the channel weights of different sensor modalities for 'rope jumping'

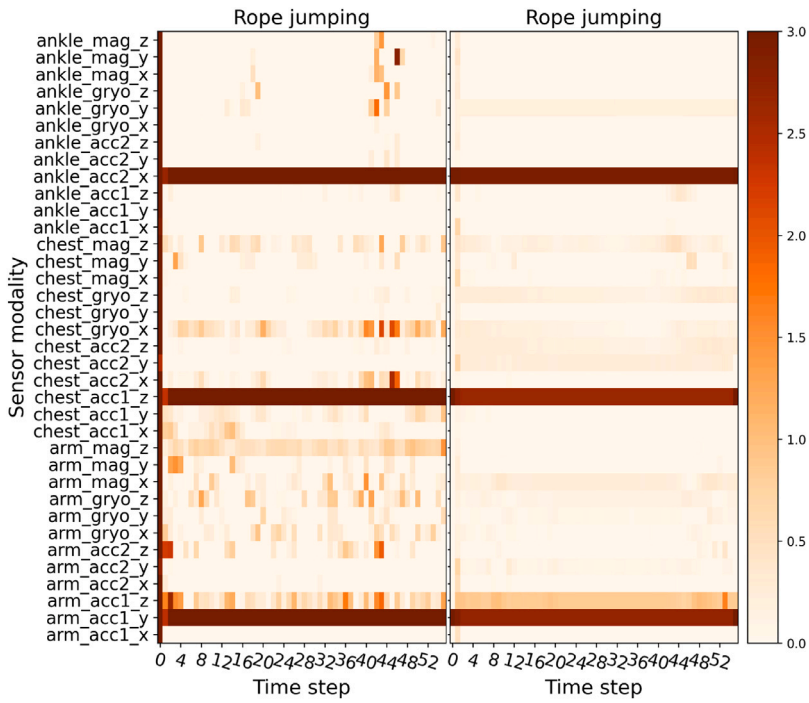


Fig. 12. Visualization of channel attention of 'rope jumping'. The left picture is base model's and the right figure is based on heterogeneous convolution.

activity. Compared with baseline, it is very clear that the proposed approach can put higher emphasis on the ankle sensor (Acc2), chest sensor (Acc1), and the arm sensor (Acc1), which is more reasonable.

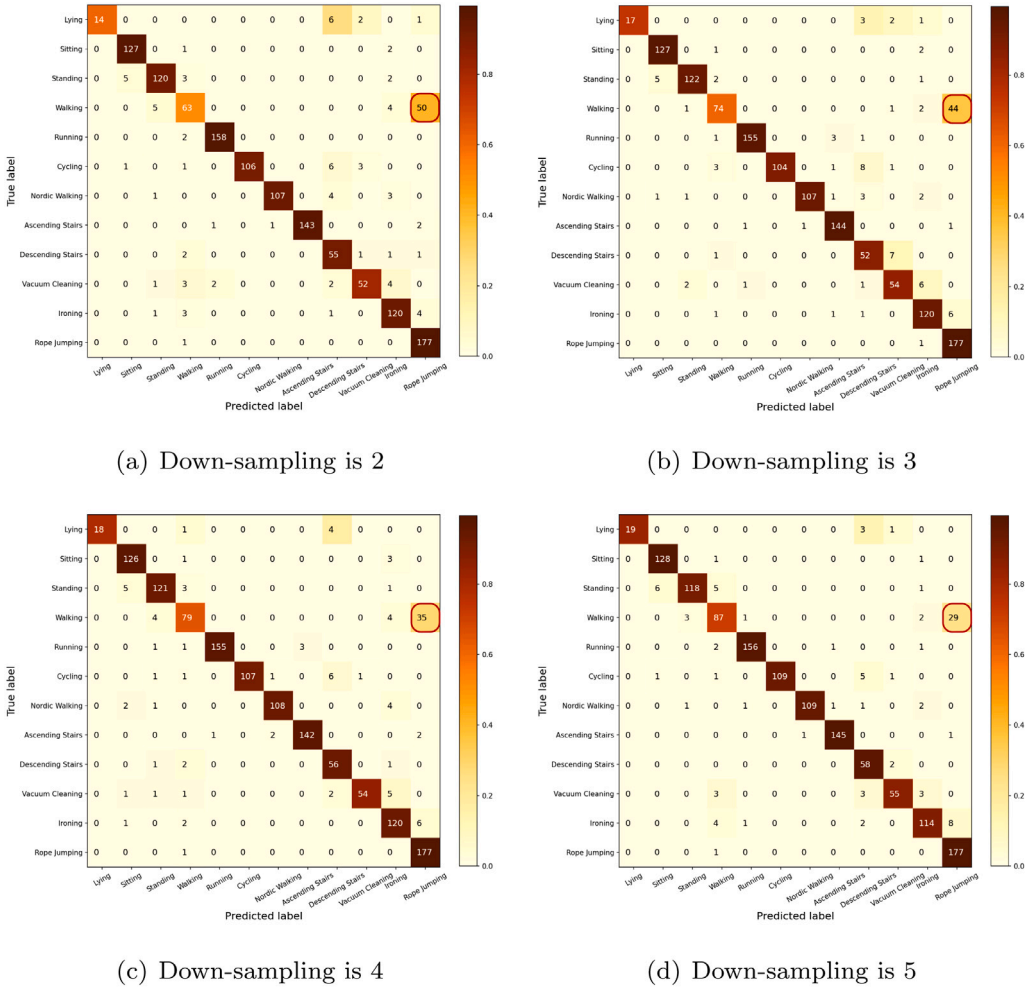
We visually show the confusion matrices of the heterogeneous two-stream CNN. Without loss of generality, the PAMAP2 dataset is chosen for our evaluation, which is commonly recognized as a realistic and challenging activity recognition task. Results are shown in Fig. 13. The x-axes represent the predicted activity samples and the y-axes represent the true activity samples, while the diagonal values represent the number of samples to be recalled. Note that the confusion between 'walking' and 'rope jumping' is very high, which is due to that the signal fluctuations between them are very similar. When the down-sampling rate is 2, the number of misclassifications reaches 50. According to accelerometer outputs caused by different speeds of 'walking'/'rope jumping', they should be more discriminative at different time scales. In the heterogeneous two-stream CNN structure, the receptive field will enlarge as the down-sampling rate increases. When the down-sampling rate is set to 3, 4, and 5, the number of misclassifications is reduced to 44, 35, and 29 respectively. The ablation studies show that the heterogeneous two-stream convolution can effectively enhance the learning capability of vanilla convolution for activity recognition at different temporal scales, which is in good line with common intuition.

Finally, for efficient consideration, we evaluate the actual operation of the proposed method in an embedded system. Without loss of generality, the Raspberry Pi 3B plus with ARM Cortex-A53 and 1 GB SDRAM is used as our test platform, because the PyTorch library can work well on Raspberry Pi. Specifically, we train the heterogeneous CNN on UCI-HAR dataset and load this trained model into the embedded platform. We perform the timing after the model is loaded and starts to output a prediction. A Raspberry Pi-based application is developed for activity recognition, and its user interface is shown in Fig. 14. The baseline and our network take 160.86–178.29 ms and 170.35–186.26 ms respectively to predict one sample. As we can see in Section 4, there is still a small gap with respect to the theoretical FLOPs. However, when evaluating actual implementation of grouped convolutions, several researchers have verified that the measured inference times are far from the expected ones due to high memory access cost. We leave this investigation in future work. Despite this, a 2.56-s window is used to segment sensor time series, where the sliding step length is equal to 1.28 s. That is to say, the recognition system will wait for 1.28s to predict next sample. The proposed method can easily meet the runtime requirement on the resource-constrained embedded system.

6. Conclusion

In this paper, we propose a novel convolution operation for activity recognition task, which can heterogeneously exploit the convolutional filters within a specific convolutional layer. To make the filters to be more diverse, we introduce a down-sampling operation to adjust the receptive field within one filter group, which is used to recalibrate the other normal filter group. Our experiments indicate obvious advantages of heterogeneous convolution in ubiquitous HAR scenario, which can lead to significant performance gain across a wide range of HAR application domains without adjusting network's architecture. Overall, through the state-of-the-art examples, the heterogeneous two-stream convolution structure shows a great potential to recognize human activities from sensory data. Due to the heterogeneous convolution operation, the receptive field, i.e., window length for each specific activity can be adjusted, which allows CNN to encode contextual information at different temporal scales, hence making the extracted activity features more discriminative. We visually show the feature representations generated by the heterogeneous convolution with different down-sampling rates, which indicates its superiority over standard CNN. We also discuss heterogeneous convolution's efficiency and effectiveness by lots of ablation studies. The proposed heterogeneous convolutions can be easily integrated into deep models for HAR with little computational overhead, rendering our method applicable for practical HAR deployment. We hope it can encourage further study about how to heterogeneously exploit convolutional filters, which can provide the HAR research community a different perspective on activity feature extraction from raw sensory data.

HAR has been widely utilized to monitor activities of a user continuously in an unobstructive manner. As we have known, for the same activity, the performed way may potentially vary among users. Therefore, an attacker may infer discriminative user-sensitive information, e.g., identity, gender, weight, height and age from time series



(a) Down-sampling is 2

(b) Down-sampling is 3

(c) Down-sampling is 4

(d) Down-sampling is 5

Fig. 13. Confusion Matrices on PAMAP2 Dataset with Different Down-sampling Rates.

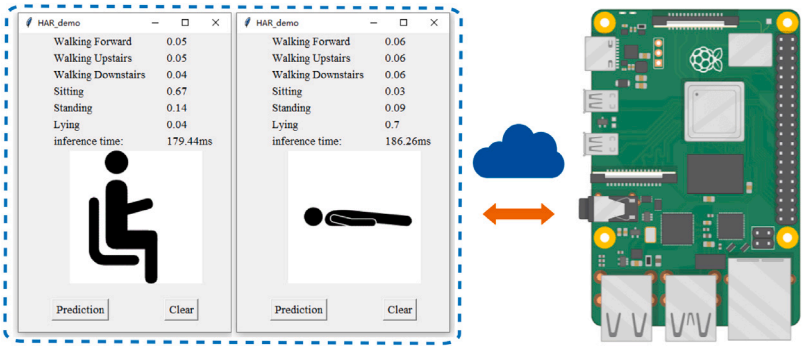


Fig. 14. The User Interface of Raspberry Pi.

sensor data. This is due to intrinsic black-box characteristic of deep learning, which will always be at the risk of revealing unintentionally user-sensitive information. Thus, it is a critical concern to deal with the privacy leakage issue of HAR by deep learning models. For example, in Iwasawa et al. (2017), the authors explore the privacy protection problem that uses discriminative features extracted by CNN for activity recognition. Their studies indicate that the CNN intentionally designed for activity classification still shows a strong ability to identify different users. To resolve this issue, one feasible strategy is to combine an adversarial loss with the conventional cross-entropy loss during training process. Specifically, the adversarial loss can be used to prevent privacy leakage by minimizing discriminative accuracy of specific private information by an end-to-end adversarial training. The cross-entropy loss function can be used to optimize the accuracy of activity recognition. Overall, the adversarial loss method has a great potential to prevent privacy leakage by reducing inferring accuracy of user-sensitive information. We will quantitatively investigate how to better tradeoff privacy protection and recognition accuracy at different time scales in a future study.

CRediT authorship contribution statement

Chaolei Han: Conceptualization, Data curation, Investigation, Methodology. **Lei Zhang:** Writing – original draft, Writing – review

& editing. **Yin Tang:** Formal analysis, Software. **Wenbo Huang:** Formal analysis, Validation. **Fuhong Min:** Methodology, Supervision. **Jun He:** Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported in part by the National Science Foundation of China under Grant 61971228 and the Industry-Academia Cooperation Innovation Fund Projection of Jiangsu Province under Grant BY2016001-02, and in part by the Natural Science Foundation of Jiangsu Province, China under grant BK20191371.

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1–6). IEEE.

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2012). Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International workshop on ambient assisted living* (pp. 216–223). Springer.

Bi, H., Perello-Nieto, M., Santos-Rodriguez, R., & Flach, P. (2020). Human activity recognition based on dynamic active learning. *IEEE Journal of Biomedical and Health Informatics*, *25*(4), 922–934.

Bruckstein, A. M., Elad, M., & Kimmel, R. (2003). Down-scaling for better transform compression. *IEEE Transactions on Image Processing*, *12*(9), 1132–1144.

Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, *46*(3), 1–33.

Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S. T., Tröster, G., Millán, J. d. R., & Roggen, D. (2013). The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, *34*(15), 2033–2042.

Chen, Y., & Xue, Y. (2015). A deep learning approach to human activity recognition based on single accelerometer. In *2015 IEEE international conference on systems, man, and cybernetics* (pp. 1488–1492). IEEE.

Chen, K., Yao, L., Zhang, D., Guo, B., & Yu, Z. (2019). Multi-agent attentional activity recognition. In *IJCAI* (pp. 1344–1350).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, *108*, Article 107561.

Dong, Y., Li, X., Dezert, J., Zhou, R., Zhu, C., Wei, L., & Ge, S. S. (2021). Evidential reasoning with hesitant fuzzy belief structures for human activity recognition. *IEEE Transactions on Fuzzy Systems*.

Feng, S., Setoodeh, P., & Haykin, S. (2017). Smart home: Cognitive interactive people-centric Internet of Things. *IEEE Communications Magazine*, *55*(2), 34–39.

Gao, W., Zhang, L., Huang, W., Min, F., He, J., & Song, A. (2021). Deep neural networks for sensor-based human activity recognition using selective kernel convolution. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–13.

Gribbon, K. T., & Bailey, D. G. (2004). A novel approach to real-time bilinear interpolation. In *Proceedings. DELTA 2004. Second IEEE international workshop on electronic design, test and applications* (pp. 126–131). IEEE.

Hammerla, N. Y., Halloran, S., & Plötz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 1533–1540).

Haresamudram, H., Beedu, A., Agrawal, V., Grady, P. L., Essa, I., Hoffman, J., & Plötz, T. (2020). Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 international symposium on wearable computers* (pp. 45–49).

Haresamudram, H., Essa, I., & Plötz, T. (2021). Contrastive predictive coding for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(2), 1–26.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645). Springer.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, C., Chen, Y., Hu, L., & Peng, X. (2018). A novel random forests based class incremental learning method for activity recognition. *Pattern Recognition*, *78*, 277–290.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, W., Zhang, L., Gao, W., Min, F., & He, J. (2021). Shallow convolutional neural networks for human activity recognition using wearable sensors. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11.

Ignatov, A. (2018). Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, *62*, 915–922.

Iwasawa, Y., Nakayama, K., Yairi, I., & Matsuo, Y. (2017). Privacy issues regarding the application of DNNs to activity-recognition using wearables and its countermeasures by use of adversarial training. In *IJCAI* (pp. 1930–1936).

Jiang, W., & Yin, Z. (2015). Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 1307–1310).

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.

Kawaguchi, N., Ogawa, N., Iwasaki, Y., Kaji, K., Terada, T., Murao, K., Inoue, S., Kawahara, Y., Sumi, Y., & Nishio, N. (2011). HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings. In *Proceedings of the 2nd augmented human international conference* (pp. 1–5).

Khan, Z. N., & Ahmad, J. (2021). Attention induced multi-head convolutional neural network for human activity recognition. *Applied Soft Computing*, *110*, Article 107671.

Kim, P. (2017). Convolutional neural network. In *MATLAB deep learning* (pp. 121–147). Springer.

Kim, E. (2020). Interpretable and accurate convolutional neural networks for human activity recognition. *IEEE Transactions on Industrial Informatics*, *16*(11), 7190–7198.

Kirkland, E. J. (2010). Bilinear interpolation. In *Advanced computing in electron microscopy* (pp. 261–263). Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kwon, H., Abowd, G. D., & Plötz, T. (2018). Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables. In *Proceedings of the 2018 ACM international symposium on wearable computers* (pp. 72–75).

Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, *15*(3), 1192–1209.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, S.-M., Yoon, S. M., & Cho, H. (2017). Human activity recognition from accelerometer data using convolutional neural network. In *2017 IEEE international conference on big data and smart computing (Bigcomp)* (pp. 131–134). IEEE.

Li, C., Niu, D., Jiang, B., Zuo, X., & Yang, J. (2021). Meta-HAR: Federated representation learning for human activity recognition. In *Proceedings of the web conference 2021* (pp. 912–922).

Liu, J.-J., Hou, Q., Cheng, M.-M., Wang, C., & Feng, J. (2020). Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10096–10105).

Luo, F., Khan, S., Huang, Y., & Wu, K. (2021). Binarized neural network for edge intelligence of sensor-based human activity recognition. *IEEE Transactions on Mobile Computing*.

Ma, H., Li, W., Zhang, X., Gao, S., & Lu, S. (2019). AttnSense: Multi-level attention mechanism for multimodal human activity recognition. In *IJCAI* (pp. 3109–3115).

Mastyło, M. (2013). Bilinear interpolation theorems and applications. *Journal of Functional Analysis*, *265*(2), 185–207.

Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, *105*, 233–261.

Ogbuabor, G., & La, R. (2018). Human activity recognition for healthcare using smartphones. In *Proceedings of the 2018 10th international conference on machine learning and computing* (pp. 41–46).

Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, *16*(1), 115.

Reiss, A., & Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers* (pp. 108–109). IEEE.

Ronao, C. A., & Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, *59*, 235–244.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, *60*(1–4), 259–268.

Sibi, P., Jones, S. A., & Siddarth, P. (2013). Analysis of different activation functions using back propagation neural networks. *Journal of Theoretical and Applied Information Technology*, *47*(3), 1264–1268.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sun, M., Song, Z., Jiang, X., Pan, J., & Pang, Y. (2017). Learning pooling for convolutional neural network. *Neurocomputing*, *224*, 96–104.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Wang, Y., Cang, S., & Yu, H. (2019). A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, *137*, 167–190.

Wang, K., He, J., & Zhang, L. (2019). Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. *IEEE Sensors Journal*, *19*(17), 7598–7604.

Wang, K., He, J., & Zhang, L. (2021). Sequential weakly labeled multiactivity localization and recognition on wearable sensors using recurrent attention networks. *IEEE Transactions on Human-Machine Systems*.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).

Xi, R., Hou, M., Fu, M., Qu, H., & Liu, D. (2018). Deep dilated convolution on multimodality time series for human activity recognition. In *2018 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Xia, S., Chu, L., Pei, L., Zhang, Z., Yu, W., & Qiu, R. C. (2021). Learning disentangled representation for mixed-reality human activity recognition with a single IMU sensor. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–14.

Xiao, Z., Xu, X., Xing, H., Song, F., Wang, X., & Zhao, B. (2021). A federated learning system with enhanced feature extraction for human activity recognition. *Knowledge-Based Systems*, *229*, Article 107338.

Yang, J., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-fourth international joint conference on artificial intelligence*.

Zeng, M., Gao, H., Yu, T., Mengshoel, O. J., Langseth, H., Lane, I., & Liu, X. (2018). 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).

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., & Zhang, J. (2014). Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services* (pp. 197–205). IEEE.

Zhang, M., & Sawchuk, A. A. (2012). USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 1036–1043).

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848–6856).

Zhou, X., Liang, W., Kevin, I., Wang, K., Wang, H., Yang, L. T., & Jin, Q. (2020). Deep-learning-enhanced human activity recognition for internet of healthcare things. *IEEE Internet of Things Journal*, *7*(7), 6429–6438.