Deep Neural Networks for Sensor-Based Human Activity Recognition Using Selective Kernel Convolution

Wenbin Gao, Lei Zhang, Wenbo Huang, Fuhong Min, Jun He, and Aiguo Song, *Senior Member, IEEE*

***Abstract*— Recently, the state-of-the-art performance in various sensor-based human activity recognition (HAR) tasks has been acquired by deep learning, which can extract automatically features from raw data. In standard convolutional neural networks (CNNs), there is usually the same receptive field (RF) size of artificial neurons within each feature layer. It is well known that the RF size of neurons is able to change adaptively according to the stimulus, which has rarely been exploited in HAR. In this article, a new multibranch CNN is introduced, which utilizes a selective kernel mechanism for HAR. To the best of our knowledge, it is for the first time to adopt an attention idea to perform kernel selection among multiple branches with different RFs in the HAR scenario. We perform extensive experiments on several benchmark HAR datasets, namely, UCI-HAR, UNIMIB SHAR, WISDM, PAMAP2, and OPPORTUNITY, as well as weakly labeled datasets. Ablation experiments show that the selective kernel convolution can adaptively choose an appropriate RF size among multiple branches for classifying numerous human activities. As a result, it can achieve a higher recognition accuracy under a similar computing budget.**

***Index Terms*— Attention, convolutional neural network (CNN), human activity recognition (HAR), kernel selection, sensor.**

I. Introduction

With the continuous technological advancement, ubiquitous sensing, which aims to extract knowledge from raw data acquired by pervasive sensors, has become a new research hotspot [1]. In particular, the current sensing devices, such as fitness trackers, smartwatches, or phones that incorporate various inertial sensors, such as accelerometers and gyroscopes, have been widely applied for analyzing numerous human activities [2]. This opens up a new research area within intelligent applications, in which traditional machine learning (ML) algorithms have been utilized to recognize simple or complex human activities. Among these popular examples are smart homes, health care, human–machine interaction, and fitness tracking applications [3]. For example, smart homes could be very beneficial for residents to enhance their living quality. In the smart home scenario, Zhang *et al.* [4] proposed a novel knowledge-based approach for multiagent cooperation, which is able to accurately recognize multiple activities, where service robots can offer proper services according to recognition results. Chen *et al.* [5] presented a novel smartphone-based human activity recognition (HAR) method that combines handcrafted features and automatic features to further boost recognition accuracy in HAR. On the whole, the HAR that uses sensor time series generated by inertial sensors embedded into wearable devices, such as smartphones, has become one dominant technique due to its obvious advantage compared with other sensor modalities, such as cameras.

Classical ML algorithms, i.e., KNN, SVM, and ensemble learning, have achieved appealing results in inferring human activities [6]. However, these shallow learning methods have to heavily rely on handcrafted features, which often requires specific expert knowledge. Recently, convolutional neural networks (CNNs) have become one dominant HAR technique, which can alleviate the manually designed burden and automatically learn features. Even though CNNs have significantly surpassed these classical ML algorithms with handcrafted features, there are still some challenges in the ubiquitous HAR scenario. For mainstream CNN architectures for HAR applications, the receptive fields (RFs) within each feature layer often share the same size, which is hard to collect multiscale features from various human activities.

In the neuroscience community, it is well known that, for visual cortical neurons, there are different RF sizes within the same area, which enables the neurons to collect information at multiple scales. This mechanism has been exploited in the computer vision field. For instance, inside an "inception" building unit, the multiscale features can be aggregated from multiple kernels with different filter sizes via a simple concatenation. Unfortunately, some other RF properties of cortical neurons, such as adaptive changing of RF size, have not received much attention in the CNN design. Recent studies have indicated that the RFs of the neurons within the visual cortex area can be adaptively modulated according to the stimulus. The contrast of the stimulus has a potential influence on the RF size: the smaller contrast usually leads to a larger

Manuscript received April 22, 2021; revised July 22, 2021; accepted July 26, 2021. Date of publication August 5, 2021; date of current version August 13, 2021. This work was supported in part by the National Science Foundation of China under Grant 61203237, in part by 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 under Grant BK20191371. The Associate Editor coordinating the review process was He Wen. *(Corresponding author: Lei Zhang.)*

Wenbin Gao, Lei Zhang, Wenbo Huang, and Fuhong Min are with the School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing 210023, China (e-mail: leizhang@njnu.edu.cn).

Jun He is with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Aiguo Song is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China.

Digital Object Identifier 10.1109/TIM.2021.3102735

RF size. For existing models, such as InceptionNet [7]–[9], in which there are multiple kernels with different sizes within one feature layer, the RF size can be adjusted adaptatively according to the contents of the input. However, it only aggregates multiscale information linearly from different branches, which inevitably limits the adaptation ability of neurons. Recently, several nonlinear approaches [10], [11] have been proposed to learn multiscale features from multiple paths for realizing an adaptive RF. Up until now, this mechanism has rarely been exploited in the HAR scenario. How to collect multiscale information to classify various human activities deserves deep studies.

In this article, we, for the first time, present an attention approach to learn multiscale features among multiple kernels in the HAR scenario. The adaptive RF is realized via introducing a selection kernel (SK) strategy, which is composed of three main parts: $Split$, $Fuse$, and $Select$. First, the $Split$ part is in charge of generating multiple paths that have different RF sizes. Second, in order to obtain a final representation for kernel selection, the $Fuse$ part plays an important role in collecting multiscale features from multiple paths. Finally, an attention mechanism is used in the $Select$ part to generate selection weights. Extensive experiments are conducted to show how the neurons in SK convolution adjust adaptively their RF sizes according to an input to capture multiscale features from various human activities. The experimental results indicate that the SK is significantly superior to previous state-of-the-art (SOTA) models that use the fixed-size kernel, which is accompanied by only a slight increase in the memory and computational burden. Comparing with existing work, our main contribution is threefold.

First, we, for the first time, propose a multibranch convolutional network in a wearable HAR scenario, where the attention idea is used to adaptively select an appropriate RF size among multiple branches to classify various human activities.

Second, within the traditional convolution layer, there is only a fixed RF size to address various human activities. As a comparison, our method can better understand human activities via using variable RF sizes. The experimental results show that there is an outstanding performance improvement for five benchmark HAR datasets that consist of UCI-HAR [12], UNIMIB SHAR [13], WISDM [14], PAMAP2 [15], and OPPORTUNITY [16] at similar computational overhead.

Third, ablation studies are provided to analyze the effect of several key factors, such as branch number, group number, and dilation rate in regard to typical challenges in wearable HAR scenarios. In particular, we show that kernel selection is more beneficial for learning automatic features via utilizing variable RF sizes in weakly supervised HAR tasks.

For previous CNNs in the wearable HAR scenario, there is only a fixed RF within each feature layer. In this article, we compute attention weights across multiple branches with different kernel sizes. Compared with previous attention methods, the SK allows multiple kernel sizes within each convolutional layer, in which the attention mechanism is used to adaptively select the optimal kernel for activity recognition. In fact, how to adaptively select an appropriate RF size plays an important role in understanding numerous activities. For example, when distinguishing the "walking" and "jumping" activities, the smaller RF is more beneficial for focusing on the "jumping" activity because the short signal of the "jumping" activity only appears in a small time interval. Instead, the larger RF is more beneficial for capturing salient features of the "walking" activity that often shows up in the entire window. Even for the simplest recognition task, perceiving sensor information from very different scales is essential to understand the human activity. Our performance improvement can be attributed to the use of adaptive RF, where attention is utilized to perform the adaptive selection mechanism among multiple kernels. In other words, our model has a self-adaptive adjustment ability according to the contextual information of the input.

We structure the rest of this article as follows. Section II reviews the recent literature in HAR and multibranch convolution. Section III gives an overview of the proposed HAR framework. Section IV details several public HAR datasets and weakly labeled datasets, as well as experimental design. In addition, experimental comparison and analysis are provided from several aspects. Finally, our work is concluded and the future work is discussed in Section V.

II. Related Works

In this section, recent related works are reviewed from two aspects—fixed RF algorithms and dynamic RF algorithms.

A. Fixed RF Algorithms

In recent years, CNNs have made remarkable achievements on sensor-based HAR, which is obviously superior to traditional ML algorithms that use cumbersome handcrafted features. For example, in order to maintain scale invariance character within an acceleration signal, Zeng *et al.* [17] first used CNNs to learn discriminative features for recognizing various human activities. Yang *et al.* [18] adopted CNNs to learn automatic features from raw time series, which can be represented from low to high level in a hierarchical way. In order to alleviate the burden of sensor data annotation, Wang *et al.* [19] presented a soft attention method to classify human activities in the weakly supervised learning scenario. Hammerla *et al.* [20] rigorously explored deep convolutional and recurrent networks across thousands of recognition experiments on various benchmark HAR datasets to investigate the applicability of each model for different HAR tasks. By assembling the time sequence of accelerometer and gyroscope into a 2-D activity image, Jiang and Yin [21] proposed a novel CNN that is able to extract optimal features from the image for the use of HAR. In the multimodal HAR scenario, Ordóñez and Roggen [22] presented a DeepConvLSTM network, which combines convolutional and LSTM units for fusing multimodal sensor information to improve recognition performance. Hu *et al.* [23] presented a novel random forests method that utilizes class incremental learning method for activity recognition. Qian *et al.* [24] ensembled convolutional and recurrent models into a unified framework to automatically learn useful temporal features, statistical features, and spatial

correlation features, which are then concatenated into one feature map for the final activity recognition. Teng *et al.* [25] used local error signals to train CNN layer by layer, which can produce higher classification performance for HAR at a much lower memory budget. Li *et al.* [26] proposed a general evaluation framework for HAR, which allows a rigorous comparison of features learning by various approaches. Tang *et al.* [27] introduced a layerwise training CNN with smaller Lego filters for HAR using wearable sensors, which can greatly reduce model complexity. Ronao and Cho [28] presented deep CNN to perform efficient and effective HAR using smartphone sensors, which provides a way to automatically extract features from raw time-series signals. Long *et al.* [29] designed an asymmetric residual neural network for accurate HAR task. Ignatov [30] used a CNN that combines automatic features and handcrafted features to perform real-time HAR tasks from raw acceleration signals. Zeng *et al.* [31] embedded two continuous attention modules into recurrent networks along temporal and sensor axes, respectively, to improve the understandability of HAR. In the semisupervised scenario, Alsheikh *et al.* [32] presented deep learning models to handle HAR tasks using spectrogram signals instead of raw acceleration data. Khan *et al.* [33] proposed a useful approach to optimize sampling frequencies of acceleration signals, which can effectively tailor activity inference systems according to particular scenarios.

Recent studies have indicated that the RF sizes of the neurons within the visual cortical area can change adaptively according to the stimulus. The mechanism has not received much attention in understanding various human activities. In existing CNN architectures for HAR applications, there is usually a fixed RF size within the same feature layer for artificial neurons, which prevents the neurons to collect multiscale features at the same processing stage. Thus, it deserves deep studies into lightweight CNN architectures with multiple kernels at different scales.

B. Dynamic RF Algorithms

From the biological viewpoint, the RFs of visual cortical neurons can vary adaptively according to the stimulus. At the end of the last century, several researchers have found that the sizes of classical RFs and nonclassical RFs can be adjusted adaptively via changing the contrast of the stimulus [34]–[36]. Unfortunately, this idea has not been fully exploited in the model design of deep networks. In the computer vision field, one important strategy is multibranch convolution [37]. Highway networks [38] were first proposed, which introduce bypassing paths via using a gating unit. Within the two-branch architecture, the highway networks make train deep networks with hundreds of layers feasible. This strategy is further inherited by ResNet [39], in which the bypassing path is used as an identity mapping. The BlockDrop [40] used more identical paths to construct major transformation. In the InceptionNet [7]–[9], multiple branches are linearly combined with customized kernel filters, which can aggregate more informative and multifarious features. Recently, attention has been used to fuse multiple kernels within the same convolutional layer, which is able to yield the effective RF with different sizes in the fusion layer [10], [11], [41]–[43].

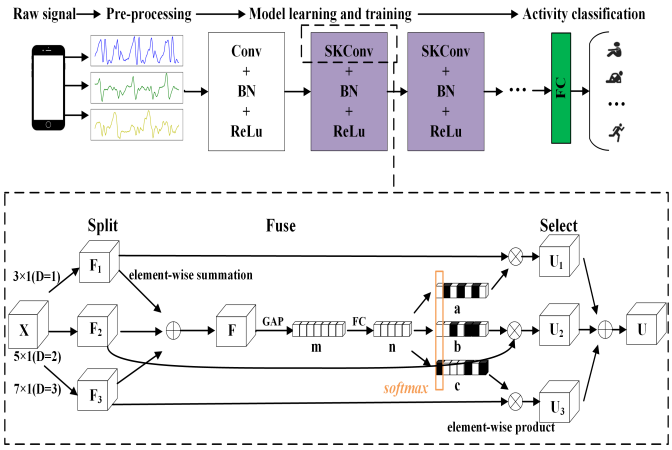


Fig. 1. Overview of the proposed SK network. The "SKConv" means the selective kernel convolution.

Despite the success of multiple RFs' convolutions, their primary use lies in visual recognition tasks. Although multibranch convolutional networks have been extensively studied, how to select adaptively an optimal RF size to capture various human activities has not received much attention. In particular, attention mechanism has rarely been considered for the selection process in the wearable HAR scenario. In this article, we first propose a new CNN that has multiple RF sizes within the same feature layer, where attention is utilized to adaptively select an appropriate RF size to recognize various activities according to the content of an input.

III. Model

In this part, the kernel selection idea is introduced in the HAR scenario, where multiple convolution kernels are used to generate more robust features that enable each neuron to adaptively choose an appropriate RF size according to the content of the input. The SK convolution is used as the basic unit to build an efficient CNN, which is able to aggregate multiscale information from multiple kernels with different sizes, e.g., 3×1, 5×1, and 7×1. Instead of linear concatenation, the softmax attention mechanism is used in the SK convolution to fuse multiple branches with different kernel sizes. The effective RF size can be determined according to attention weights over multiple branches.

A. Efficient Implementation of SK Module

The soft attention idea is used to perform an automatic selection operation among multiple branches with different kernel sizes. The SK convolution is used to replace standard convolution, which can adaptively adjust the RF size during the convolution process. As can be seen in Fig. 1, a three-branch case is presented, but two or more branches are also feasible.

1) Split: For one feature map $\mathbf{X} \in \mathbb{R}^{C' \times H' \times W'}$, in which C' denotes channel number, H' denotes height (sensor axes), and W' denotes width (temporal axes), we first split $\mathbf{X}$

into three transformations $\mathbf{F}_1 \in \mathbb{R}^{C\times H\times W}$, $\mathbf{F}_2 \in \mathbb{R}^{C\times H\times W}$, and $\mathbf{F}_3 \in \mathbb{R}^{C\times H\times W}$ with different kernel sizes 3×1, 5×1, and 7×1, respectively. In this way, the three transformations, i.e., $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$, consist of grouped convolutions [10], [44], batch normalization (BN) [45], and ReLU activation [46] in sequence. The group number G is an important factor in grouped convolution, which has first been proposed in AlexNet [47]. In comparison with ordinary convolution, the parameter number and computational burden of the model are divided into G parts, in order to distribute the deep model over more GPU resources. In our design, both grouped convolution [10], [44] and dilated convolution [48] have been integrated into the branches with larger kernel size, which can avoid heavy model overheads. The dilation D is a hyperparameter called the dilation rate in dilated convolution. For convolutional networks, it is an important way to expand the receptive view. The standard convolution with the kernel size of 5×1 can be replaced by dilated convolution with the kernel size of 3×1 and the dilation size of 2. The standard convolution with the kernel size of 7×1 can be replaced by dilated convolution with the kernel size of 3×1 and the dilation size of 3 and so on. As we have known, the dilated convolution has lower model complexity. In short, dilated convolution is a simple but effective idea, which can be widely used in both cases: a broader view of the input to capture more contextual information for human activities and faster run-time with fewer parameters.

2) Fuse: We fuse results from all three branches via implementing an elementwise summation

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3. \tag{1}$$

Recent studies have indicated that global average pooling (GAP) [10], [49]–[51] can be used to generate channelwise information. In order to aggregate features more effectively, we feed the feature $\mathbf{F}$ through a simple GAP to infer the channelwise statistics as $\mathbf{m} \in \mathbb{R}^C$, and the cth element of $\mathbf{m}$ can be formulated as

$$\mathbf{m}_c = \text{AvgPool}(\mathbf{F}_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{F}_c(i, j). \tag{2}$$

Subsequently, a fully connected (FC) layer is used to generate a more compact feature map $\mathbf{n} \in \mathbb{R}^{d\times 1}$, which enables the model to perform the SK operation in an efficient way. During the dimension reduction, the value of d is controlled by a reduction ratio r. The compact feature map $\mathbf{n}$ and the reduction ratio r can be computed as

$$\mathbf{n} = \text{FC}(\mathbf{m}) = \text{ReLU}(\beta(\mathbf{W}\mathbf{m})) \tag{3}$$

$$d = C/r \tag{4}$$

where β refers to the BN operation, $\mathbf{W} \in \mathbb{R}^{d\times C}$, and r represents the reduction ratio of the compact feature map $\mathbf{n}$.

3) Select: In this part, we apply soft attention mechanism for the compact feature map $\mathbf{n}$ passed down from the previous layer, which is able to guide the model to adaptively extract multiscale information across the channel axis [10], [11], [19]. The softmax attention, that can focus on the important branches, plays a key role in the adaptive kernel selection

$$a_c = \frac{e^{\mathbf{A}_c\mathbf{n}}}{e^{\mathbf{A}_c\mathbf{n}} + e^{\mathbf{B}_c\mathbf{n}} + e^{\mathbf{C}_c\mathbf{n}}} \tag{5}$$

$$b_c = \frac{e^{\mathbf{B}_c\mathbf{n}}}{e^{\mathbf{A}_c\mathbf{n}} + e^{\mathbf{B}_c\mathbf{n}} + e^{\mathbf{C}_c\mathbf{n}}} \tag{6}$$

$$c_c = \frac{e^{\mathbf{C}_c\mathbf{n}}}{e^{\mathbf{A}_c\mathbf{n}} + e^{\mathbf{B}_c\mathbf{n}} + e^{\mathbf{C}_c\mathbf{n}}} \tag{7}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{C\times d}$, and $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are the soft attention feature maps generated from the split feature maps $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$. Note that $\mathbf{A}_c \in \mathbb{R}^{1\times d}$ is the cth row of $\mathbf{A}$, and a_c is the cth element of $\mathbf{a}$, likewise $\mathbf{B}_c$, b_c, $\mathbf{C}_c$, and c_c. The final feature map $\mathbf{U}$ is superimposed by the attention weights with different convolution kernels on all branches

$$\mathbf{U}_c = a_c \cdot \mathbf{F}_{1c} + b_c \cdot \mathbf{F}_{2c} + c_c \cdot \mathbf{F}_{3c}, \quad a_c + b_c + c_c = 1 \tag{8}$$

where $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \ldots, \mathbf{U}_c]$, where $\mathbf{U}_c \in \mathbb{R}^{H\times W}$. Note that the deduction process can easily extend to other cases when the branches are two or more.

B. Efficient Implementation of SK Networks

Multiple SK units can be stacked to build an SK network. By replacing standard convolution with SK convolution, we are able to achieve very compelling results in HAR. In this article, the shorthand description of the SK network is Conv64-SKConv128-SKConv256-FC-Softmax. As mentioned above, there are three important hyperparameters that affect performance: the branch number M that determines how many convolution kernels can be selected, and the group number G and dilation D that control the cardinality of each branch. The influence caused by their changes will be discussed in Section IV. Ablation studies are conducted to find the optimal settings via analyzing the independent contribution of each part.

IV. Experiment

At the preprocessing stage, it is an important step to split sensor time series into a series of windows for activity recognition. The sliding window with a fixed window length and overlap has been extensively leveraged to perform segmentation. As a result, the streams of sensor data are often divided into continuous windows, where each window may be assigned a specific activity label. For a large variety of HAR tasks, there is still no clear consensus on which is the optimal window length to be preferably selected. How does sliding window affects the activity recognition accuracy still remains unclear. So far, for what is the optimal window size across a large variety of HAR tasks, this challenge has been rarely and vaguely addressed. Online activity recognition focuses on collecting data and recognizing activities in real time, which is necessary if one user requires instantaneous feedback from the system, such as monitoring higher risk patients. There is great demand for online activity inference, where such a recognition system can track the execution of the real-time activity. However, for online recognition, it has to wait for future data to make the decision. According to common intuition, for a longer window, the recognition system has to

TABLE I
SIMPLE DESCRIPTION OF DATASETS

Attribute \ Dataset	UCI-HAR	UNIMIB SHAR	WISDM	PAMAP2	OPPORTUNITY	Weakly Labeled
Sampling Rates	50Hz	50Hz	20Hz	33.3Hz	30Hz	50Hz
Number of Categories	6	17	6	18	18	4
Sliding Window Size	2.56s	3s	10s	5.12s	1s	40.96s
Overlap Rates	50%	50%	95%	78%	50%	50%
Subjects	30	30	29	9	4	7
Proportion of Training Data	70%	70%	70%	80%	70%	70%
Proportion of Testing Data	30%	30%	30%	20%	30%	30%

TABLE II
SIMPLE DESCRIPTION OF THE BASELINE CNN AND SK NETWORKS

Simple Description \ Dataset	UCI-HAR	UNIMIB SHAR	WISDM	PAMAP2	OPPORTUNITY	Weakly Labeled
Layer1	C(64)	C(64)	C(64)	C(64)	C(64)	C(32)
Layer2	C(128)/SK(128)	C(128)/SK(128)	C(128)/SK(128)	C(128)/SK(128)	C(128)/SK(128)	C(64)/SK(64)
Layer3	C(256)/SK(256)	C(256)/SK(256)	C(256)/SK(256)	C(256)/SK(256)	C(256)/SK(256)	C(128)/SK(128)
FC	✓	✓	✓	✓	✓	✓
Softmax	✓	✓	✓	✓	✓	✓
Training time(epoch)	500	500	500	500	500	500
Batch size	256	200	210	256	200	300
Learning rate	0.001	0.001	0.001	0.001	0.001	0.001

"wait" a longer time period for a new window to be available for classifying. It is more beneficial to reduce window length for faster activity detection, which is accompanied by reduced computing resources. Instead, raising window length is normally designed for the recognition of complex activities, which occurs for a long time period. A tradeoff between inference time and recognition performance should be considered by the recognition system. Most existing designs basically depend on the window length used in the previous literature. To proceed with fair comparison, we select the same window length in previous successful cases, in which the specific window length and overlap on each dataset are shown in Table I [12]–[16].

In addition, we perform the same preprocessing technique, which has been well established in several benchmark HAR datasets. For example, a Butterworth low-pass filter with a cutoff frequency is adopted to remove the gravitational component from acceleration signals. Sensor data at the beginning and the end of each labeled sample are removed in order to avoid handling eventual transient activities. Missing sensor values can be processed by either linear interpolation or repetition of previous values. By subtracting the mean and dividing by standard deviation, the heterogeneous sensor values are normalized into zero mean and unit variance.

We divide the whole experiment into three parts. First, in order to verify the efficiency of the SK method, extensive experiments are performed on five benchmark datasets, consisting of UCI-HAR [12], UNIMIB SHAR [13], WISDM [14], PAMAP2 [15], and OPPORTUNITY [16] in the supervised HAR scenario. Second, we analyze the effect of the SK method in the weakly supervised learning scenario. Third, we do a series of ablation experiments to analyze the attention weights across multiple branches within the same layer via changing the three crucial elements mentioned above.

One three-layer CNN is built as our baseline, in which the SK convolution is used to replace standard convolution at the second and third layers. Recently, squeeze-and-excitation networks (SENets) [51] have introduced an effective, lightweight mechanism to recalibrate the feature map via channelwise importance, which has achieved SOTA performance. We also compare our method with SENet [51]. The batch size and the initial learning rate are shown in Table II. The learning rate is set to decay exponentially. The Adam optimization method and BN are used to train our model.

A. Experiment Results and Performance Comparison

In Table I, we summarize various attributes of the five public HAR datasets and the weakly labeled dataset. The sliding window method was used to process datasets. Table II illustrates the shorthand structures of the model design. Experimental results are demonstrated in Table III, which includes a comprehensive list of the recognition accuracy obtained from past published SOTA methods.

1) UCI-HAR Dataset [12]: In order to test various ML algorithms in the HAR scenario, the researchers in the University of California at Irvine (UC Irvine) recruited 30 subjects to collect this dataset. All subjects whose age is from 19 to 48 years were asked to wear a smartphone (Samsung Galaxy S II) on their waists. One three-axis accelerator embedded into the smartphone was used to generate sensor time series, which is sampled at a frequency of 50 Hz. This dataset contains six different kinds of ADLs, such as "sitting,"

TABLE III
ACCURACY (%) PERFORMANCE OF MODELS ON VARIOUS DATASETS

Model + Method \ Dataset	UCI-HAR	UNIMIB SHAR	WISDM	PAMAP2	OPPORTUNITY	Weakly Labeled
Baseline	96.11&0.37M	74.83&0.42M	96.83&0.29M	91.21&0.76M	90.86&90.60*&0.49M	89.86&0.20M
Baseline + SE blocks	96.60&0.39M	75.56&0.43M	97.42&0.30M	92.68&0.77M	91.84&91.70*&0.51M	90.84&0.21M
Baseline + SK convolution	**97.21&0.45M**	**76.84&0.54M**	**98.13&0.36M**	**93.54&0.85M**	**93.06&93.01*&0.64M**	**92.85&0.29M**
Other Researchers' Results	**96.98** [25]	74.66 [26]	97.50 [25]	**93.03** [25]	92.70*[20]	**90.04** [19]
	96.90 [27]	74.46 [27]	96.90 [27]	86.00 [33]	89.15 [23]	-
	95.75 [28]	**76.04** [29]	93.32 [30]	89.96 [31]	**92.70***[24]	-
	95.18 [21]	-	**97.51** [32]	-	-	-
	96.37 [30]	-	-	-	-	-

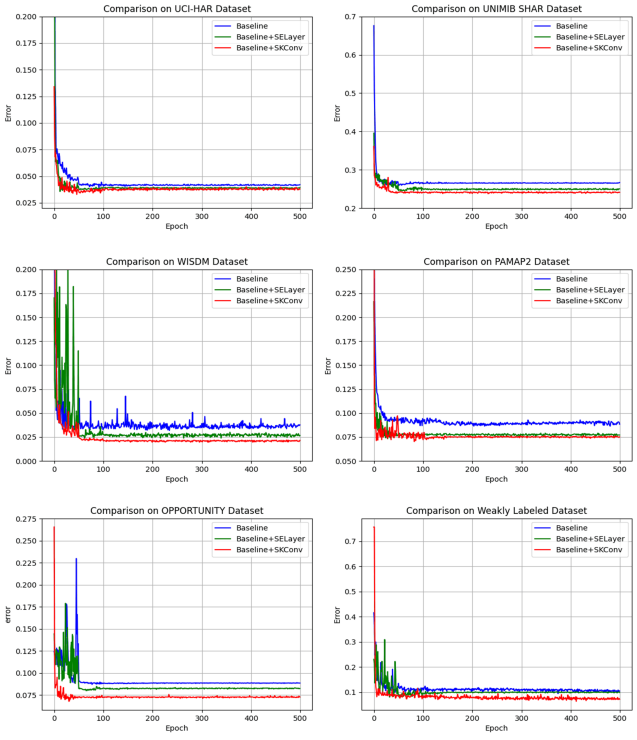


Fig. 2. Test errors of different models on multiple public datasets.

"lying," "standing," "walking upstairs," "walking downstairs," and "walking."

The SK model is first compared with baseline and SENet. According to a test error, what we can see in Fig. 2 is that the SK performs the best among the three algorithms, which is able to yield 1.10% and 0.61% accuracy gains with negligible computational overhead. We also compare our accuracy with previous SOTA results that use fixed kernels. Performance comparisons are listed in Table III. To the best of our knowledge, the existing SOTA result on the UCI-HAR dataset is 96.98% [25]. Our result obtained from the SK method is best reported, which outperforms the SOTA result with an 0.23% accuracy improvement.

2) UNIMIB SHAR Dataset [13]: This dataset was collected by the researchers from the University of Milano-Bicocca. This new dataset is designed for detecting various "falling" activities. A smartphone with the Android operation system was adopted to collect data from 30 subjects whose ages range from 18 to 60 years. All subjects participating in the data collection were asked to wear smartphones in their right and left pockets. The sampling rate of the sensor signal is 50 Hz.

Fig. 2 displays the test error curves on the UNIMIB SHAR dataset, which indicates that the SK model is superior to baseline, as well as SENet. We also compare our recognition accuracy with several previously published results that use deep learning techniques, and performance comparisons are shown in Table III. The highest reported accuracy to our knowledge is 76.04% using dual residual networks [29]. Remarkably, our model outperforms the dual residual network by above absolute 0.80% with comparable complexity.

3) WISDM Dataset [14]: In this dataset, one triaxial accelerometer sensor embedded in an Android smartphone is used to generate sensor time series. In supervised conditions, wearing the smartphone in front leg pocket, each subject carried out six common activities, namely, "walking," "jogging," "going upstairs," "going downstairs," "sitting," and "standing." The sampling rate of the sensor signal is 20 Hz. At a 95% overlapping rate, a 10-s window is shifted over sensor time series to generate samples.

In comparison with baseline and SENet, Fig. 2 indicates that the SK method clearly works better for the tested architectures, which achieves lower test errors. Table III shows that the adaptive kernel selection is able to produce 1.30% and 0.71% relative accuracy gains with negligible computational overhead. We also compare the SK method with previously published results that use convolutional networks. The best reported accuracy on the WISDM dataset, to the best of our knowledge, is 97.51% using deep CNNs [32]. Our result obtained from the SK method is the best.

4) PAMAP2 Dataset [15]: In this dataset, six subjects who wear three IMUs take part in the data collection process. The three IMUs consisting of accelerometer, gyroscope, and magnetometer are fixed to each subject's chest, wrist, and side's ankle, respectively. Each subject was asked to perform 12 protocol activities ("lying down," "standing," and so on) and six optional activities ("watching TV," "folding laundry," and so on). The sampling rate of IMUs is 100 Hz, which is further subsampled into 33.3 Hz. At a 78% overlap rate, a 5.12-s time window is utilized to slide over the sensor time series, which produces around 473k samples.

We plot the test error curves of three compared architectures in Fig. 2, which shows that the SK is consistently able to

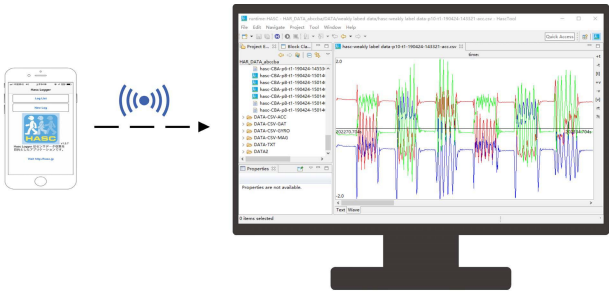


Fig. 3. UI of the acquisition software HascLogger.

Fig. 4. Data collection process of the weakly labeled dataset. The target activities from left to right are "going upstairs," "going downstairs," "jumping," and "jogging."

achieve lower test errors. The SK not only boosts the accuracy of baseline significantly but also outperforms SENet. Despite multiple paths, there is only a slight increase in computational burden. The SK convolution is compared with other SOTA algorithms. As can be seen in Table III, the SK performs the best among all the algorithms.

5) OPPORTUNITY Dataset [16]: In the dataset, each volunteer was instructed to carry out common kitchen activities at five runs. Sensor recordings with 113 dimensions are collected through inertial sensors placed on four subjects, in which the IMUs are placed at 12 on-body positions, i.e., the back and both feet. For a fair comparison, we use the same subset in the recent OPPORTUNITY challenge, where 17 mid-level human gestures are annotated and null class is also considered. The sampling rate is 30 Hz. The window with a fixed size of 1 s is moved over sensor time series with 50% overlap.

Performance comparisons on the OPPORTUNITY dataset are illustrated in Fig. 2. We can find that the SK method achieves lower test errors compared with baseline and SENet. Because the OPPORTUNITY dataset is highly imbalanced where the Null class accounts for more than 75% of the whole dataset, recent related work also used the weighted F_1 score as a primary performance metric (especially for OPPORTUNITY). In order to proceed with fair comparison, we compute the weighted F_1 score on the OPPORTUNITY dataset. We mark the results with the symbol "$*$." We compare our method with other SOTA methods in Table III. What we can see is that the recognition performance of SK is superior to those of the other SOTA methods. In terms of accuracy, our SK method surpasses Hu *et al.*'s result [23] by 3.91%. In terms of the weighted F_1 score, our SK method surpasses Qian *et al.*'s results [24] by 0.31% (the same result is also reported by Hammerla *et al.* [20]). There is only a slight increase in the number of parameters caused by multiple paths. In particular, compared with [24] that uses ensembled models, our SK model is more lightweight.

6) Weakly Labeled Dataset: This dataset was collected using an application software called HascLogger, and its user interface (UI) is presented in Fig. 3. Wearing the smartphone in the right trouser pocket, ten subjects were instructed to carry out five kinds of daily activities, such as "walking" and "jogging." Among these activities, "walking" can be deemed as the background action, while the others are interesting actions that need to be addressed. The data collection process for one specific subject is presented in Fig. 4. For one specific action, each subject performs four runs. A 2048-length window is used to divide sensor signals with a 50% overlapping rate. Each window may contain one interesting action and other interesting actions.

In the weakly supervised learning scenario, Fig. 2 shows that our SK model is superior to baseline and SENet, which is able to obtain lower test errors. We compare SK with both models on the weakly labeled dataset, which achieves 2.99% and 2.01% accuracy improvements, respectively. In our previous sensor-based HAR works [19], the attention weights are learned via a soft-attention mechanism in order to enhance the interesting activity and weaken other irrelevant background activities within each activity window. In this case, there is only one fixed kernel size for each convolutional layer. In this article, we compute attention weights across multiple branches with different kernel sizes. Compared with our previous attention methods, the SK allows multiple kernel sizes within each convolutional layer, in which the attention mechanism is used to adaptively select the optimal kernel for activity recognition. The proposed SK method surpasses our previous attention-based CNN [19] by a large margin, which achieves an accuracy improvement of 2.81% on the weakly labeled dataset.

B. Ablation Study

In this section, to better understand how the SK convolution works, we conduct ablation studies on several benchmark HAR datasets to investigate its effectiveness. In fact, there are two crucial elements, i.e., the dilation D and group number G to control the RF size. Here, we use a two-branch case to investigate their influence. In addition, we study the effect of more branches. The confusion matrices computed on the PAMAP2 dataset are also provided. In the end, we measure the inference speed of the proposed method in an Android platform.

First, we discuss the effect of dilation D and group number G. For simplicity and without loss of generality, a two-branch case is considered, in which the setting in the first branch is fixed with the 3×1 filter ($G = 32$ and $D = 1$). Under similar model complexity, the RF size in the second branch can be adjusted as follows: 1) increase the dilation D while keeping the group number G fixed and 2) simultaneously increase the dilation D and filter size.

The optimal setting in the second branch is illustrated in Table IV, in which the setting of the first branch keeps fixed. Here, the "Final kernel size" represents the approximate

TABLE IV

RESULTS OF THE SK NETWORK WITH DIFFERENT GROUP NUMBERS AND DILATION RATES ON THE **WISDM** DATASET

Model Settings	Test Acc	Params	Flops	Final Kernel Size
3×1,D=1,G=32	98.10%	0.36M	23.41M	3×1
3×1,D=1,G=64	98.04%	0.37M	23.42M	3×1
3×1,D=1,G=128	98.01%	0.38M	23.44M	3×1
3×1,D=2,G=32	**98.13%**	**0.36M**	**23.41M**	**5×1**
5×1,D=1,G=32	**98.12%**	**0.38M**	**23.44M**	**5×1**
3×1,D=3,G=32	97.94%	0.36M	23.41M	7×1

TABLE V

RESULTS OF THE SK NETWORK WITH DIFFERENT BRANCHES ON THE **WISDM** DATASET

Model Settings	Test Acc	Params
3×1	96.83%	0.29M
5×1(D=2)	96.54%	0.29M
7×1(D=3)	96.62%	0.29M
3×1+5×1(D=2)	**98.13%**	**0.36M**
3×1+7×1(D=3)	98.07%	0.36M
5×1(D=2)+7×1(D=3)	98.06%	0.36M
3×1+5×1(D=2)+7×1(D=3)	**98.15%**	**0.37M**
3×1+5×1(D=2)+7×1(D=3)+9×1(D=4)	**98.19%**	**0.39M**
3×1+5×1(D=2)+7×1(D=3)+9×1(D=4)+11×1(D=5)	97.83%	0.42M

TABLE VI

RESULTS OF THE SK NETWORK WITH DIFFERENT REDUCTION RATIOS ON THE **WISDM** DATASET

Ratio r	Test Acc	Params	Flops
8	97.99%	0.392M	26.46M
16	98.12%	0.386M	26.45M
32	98.19%	0.382M	26.45M
64	97.84%	0.377M	26.44M

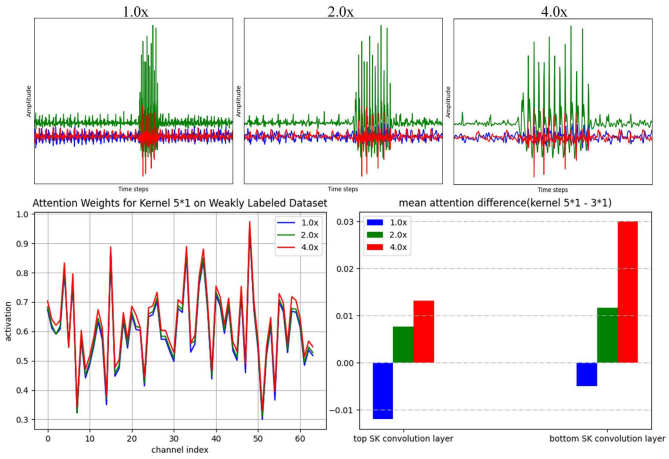


Fig. 5. Attention weights for randomly sampled "going upstairs" with three differently sized targets (1.0×, 2.0×, and 4.0×).

kernel size that is acquired from dilated convolution. It can be seen that the best result is 98.13% with 3×1, $D = 2$, and $G = 32$. The second best result is 98.12% with 5×1, $D = 1$, and $G = 32$. The result suggests that using different kernel sizes is more beneficial due to the aggregation of multiscale features. If there are the same kernel size in two branches, it may undermine the final classification results. For the two optimal configurations, i.e., the kernel size 5×1 ($D = 1$) and the kernel size 3×1 ($D = 2$), there is a slightly lower model complexity for the latter. That is to say, despite the same RF, the smaller kernel with various dilations has a significant advantage over the larger kernel without dilation in terms of performance and model complexity, which agrees well with the basic principle of the CNN design.

We next compare the performance when using more branches, in which there may be two or more kernels and their size may be larger than 3×1. Due to the limitation of search space, we only consider five-branch case, where the kernel sizes are 3×1, 5×1, 7×1, 9×1, and 11×1, respectively. Specifically, the dilated convolution is used to realize large kernels as indicated above. G is set to 32. What we can see in Table V is that the classification performance increases at first and then decreases with the increase in the branch number M. The results in the one-branch case ($M = 1$) are the worst. Due to the use of multiple kernels, the SK is able to achieve appealing results via performing the adaptive selection among multiple branches. The best result is 98.19% that corresponds to the four-branch case ($M = 4$). The second best result is 98.15% that corresponds to the three-branch case ($M = 3$). There is a negligible performance gain from $M = 3$ to $M = 4$. In the case of $M = 5$, the accuracy rapidly decreases to 97.83%. In order to better balance accuracy and efficiency, the three-branch case should be preferred.

Second, the dimension of the compact feature map $\mathbf{n}$, which is generated by the FC layer in SK convolution, can be adjusted by the reduction ratio r. In order to analyze the effect of the important hyperparameter, we conduct extensive experiments on the WISDM dataset [14] via using different r value with $M = 4$ and $G = 32$. The comparisons in Table VI reveal that the accuracy does not decrease monotonically with the increase in r. It is likely to be attributed to the reason that overfitting case occurs due to channel interdependencies during the training stage. In particular, it can be seen that the reduction ratio $r = 32$ is able to obtain the best performance in terms of accuracy.

Third, to better comprehend how selective kernel mechanism works, the attention weights of one target activity with different scales were analyzed in Fig. 5. For example, "going upstairs" can be selected as the target activity from the weakly labeled dataset, where "walking" can be seen as the background activity. The target activity can be progressively enlarged from 1.0× to 4.0× via a central cropping and subsequent resizing (top in Fig. 5). The attention weights for the large kernel 5×1 are shown (bottom left in Fig. 5). For both the kernels (5×1 and 3×1), their difference is calculated according to the mean attention weights across all channels in two SK convolution layers, respectively (bottom right in Fig. 5). The figure shows that the attention weights for the large kernel 5×1 increase when the target activity enlarges. The result suggests that the RF size within the

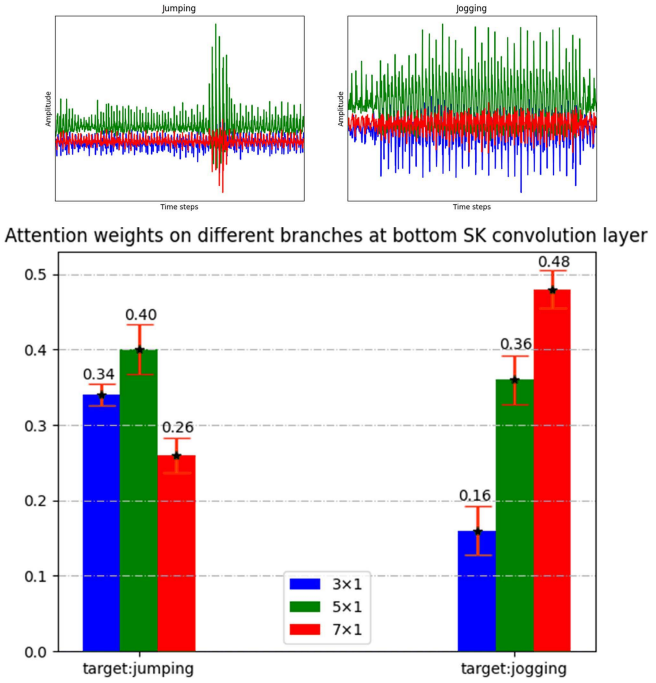


Fig. 6. Mean and standard deviation of attention weights across all test examples on three branches.

neurons in the SK model can change adaptively as the scale of the target activity varies, which agrees with our expectation.

Fourth, in order to support our conclusion, we continue to perform ablation experiments on the weakly labeled dataset. As introduced, "walking" is seen as a background activity in this dataset, which is not assigned a specific label. Without loss of generality, we select two target activities, i.e., "jumping" and "jogging" to perform a comparison. The mean and the standard deviation of attention weights across 20 test examples on three branches are illustrated in Fig. 6. For the "jumping" activity that occurs for a short time period, it can be seen that the small kernels 3×1 and 5×1 contribute more to output classification. Instead, for the "jogging" activity that takes place for a long time period (nearly entire window), we could clearly observe that the larger kernel 7×1 plays a more important role in activity recognition. The results further verify that the neurons in the proposed SK model have adaptive RF size for different activities, which is in good line with our conclusion.

Fifth, in the existing selective attention mechanisms, the SENet [51] is adopted to compute kernel attentions, in which the global feature information is first squeezed by GAP, and then, two FC layers followed by softmax are used to generate attention weights over multiple output channels. Different from SENet [51], dynamic convolution technique [11] tends to compute attention over multiple convolution kernels that share the same kernel size. Different from SENet [51] and dynamic convolution [11], the SK computes attentions over multiple branches with different kernel sizes. The advantage of the SK lies in that it can choose appropriate RF sizes in an adaptive manner. Performance comparisons show that the SK is able to consistently outperform SENet.

TABLE VII

COMPARISON OF THE SK NETWORK WITH OTHER SELECTIVE ATTENTION MECHANISMS ON THE **UCI-HAR** DATASET

General Models	Test Acc	Params	Flops
Baseline	96.11%	0.37M	13.24M
Baseline+dynamic convolution	96.27%	0.42M	14.78M
Baseline+SE blocks	96.60%	0.39M	15.11M
Baseline+SK convolution	97.21%	0.45M	17.53M

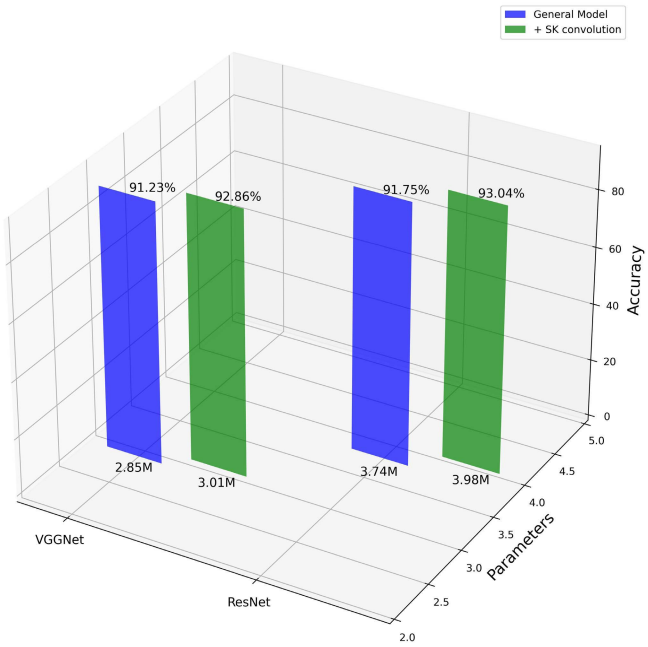


Fig. 7. Performance comparisons when SK is integrated into VGG and ResNet.

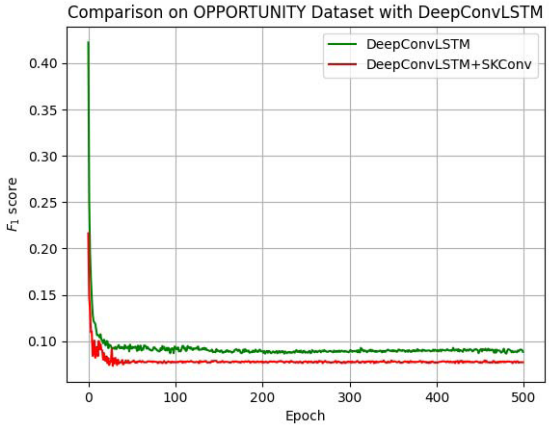
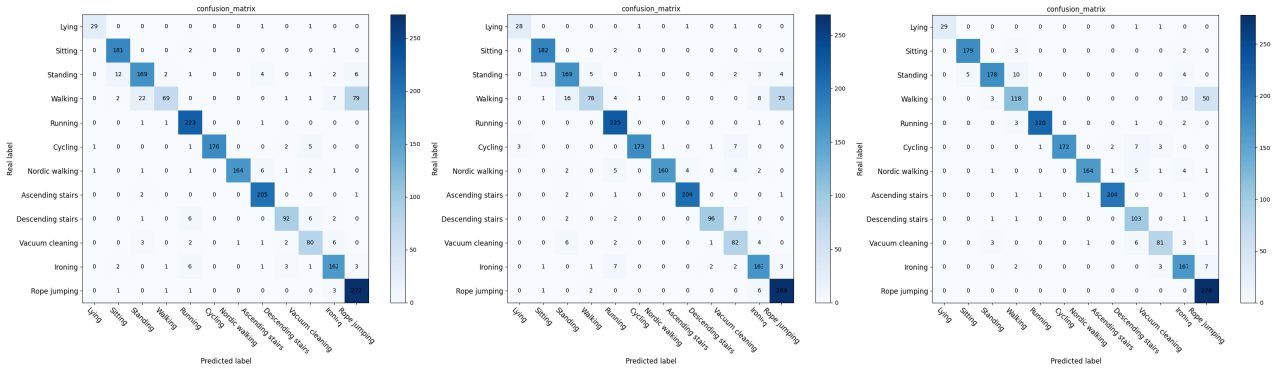


Fig. 8. Opportunity's F_1 score on DeepConvLSTM and DeepConvLSTM with SK convolution.

In addition, we also compare the SK and dynamic convolution on the UCI-HAR dataset [12]. From the results in Table VII, it can be seen that the SK can achieve more appealing results than dynamic convolution due to its adaptive RF size, which can better perceive contextual information of sensor time series for classifying human activities according to an input.



(a) (b) (c)

Fig. 9. Confusion matrix for the PAMAP2 dataset between the baseline, the baseline with SE blocks, and the SK networks from left to right. (a) Baseline. (b) Baseline + SE blocks. (c) Baseline + SK convolutions.

Sixth, the SK can be integrated into convolutional networks and their variants, such as VGG [52] and ResNet [53]. To evaluate the generality ability of our method, we further conduct ablation experiments on the OPPORTUNITY [16] dataset. We use two classic models, i.e., VGG [52] and ResNet [53], as baselines. We add the SK submodule to two baseline models in the same way as mentioned above. The comparisons between baselines and those combined with the SK of four branches ($M = 4$) are illustrated. As can be seen in Fig. 7, our method is able to achieve significantly better classification performance at almost similar computational overhead when integrated into VGG and ResNet. In order to further evaluate the generality ability of our method, we integrated the SK into the DeepConvLSTM. Ordóñez and Roggen [22] proposed to combine convolutional and LSTM layers to learn temporal features between subsequent windows, which can achieve better results than RNN or LSTM alone. We run the DeepConvLSTM code on the OPPORTUNITY [16] dataset. For a fair comparison, we select the same values used by Ordóñez and Roggen [22]. That is to say, the sensor signals are divided into smaller fixed-length windows of 500-ms duration with 50% overlap. The results in Table VIII show that our baseline DeepConvLSTM can acquire an F_1 score of 91.36%, which is very close to the result reported in the DeepConvLSTM [22]. When applying the SK method, we can further improve the F_1 score from 91.36% to 93.15%, which yields a 1.79% performance gain. The test F_1 score curves are also shown in Fig. 8. The results also indicate that smaller windows are more beneficial for the baseline DeepConvLSTM, which enables recurrent networks to better capture the temporal correlation between subsequent windows.

Seventh, we further perform comparisons via computing the confusion matrices on the PAMAP2 dataset. Due to confusion between two very similar activity classes, many of the misclassifications occur. This may be attributed to the reason that they have very similar vibrations in signal waveforms. For "rope jumping" and "walking" that were previously perceived to be very difficult to distinguish, Fig. 9 shows that baseline makes 79 errors, while SK convolution misclassifies only 50 activities, which confirms the superiority of the proposed method in recognition accuracy.

Fig. 10. Demo Application on mobile phone (Google Nexus 6).

Eighth, in order to evaluate actual inference speed of the SK network, we transfer two-branch SK model with the kernel sizes of 3×1 and 5×1 ($D = 2$) and the three-branch SK model with the kernel sizes of 3×1, 5×1 ($D = 2$), and 7×1 ($D = 3$) into an Android smartphone. Specifically, our experiment is implemented on a Google Nexus 6 phone with Android OS (11.0.0). We build an Android software to evaluate the performance, and the software's UI is illustrated in Fig. 10. The three networks are trained on the WISDM dataset [14], which are then converted into a .pb file and added as a Gradle dependence (Java) via applying into Android Studio. Someone can perform the target activity and get the classification result after loading the saved model. It can be seen from Table IX that, despite better recognition accuracy, there is only a slight increase in the inference speed.

Ninth, in order to evaluate the robustness of the proposed method, we choose the Raspberry Pi 3B plus with ARM Cortex-A53 and 1-GB SDRAM as our test platform, where the PyTorch deep learning library has good compatibility with the Raspberry PI operating system. To be specific, two main steps are implemented as follows: 1) train the network with

TABLE VIII

RESULTS OF DEEPCONVLSTM AND DEEPCONVLSTM WITH SK CONVOLUTION ON THE **OPPORTUNITY** DATASET

General Models	F_1 score	Params	Flops
DeepConvLSTM	91.36%	0.28M	6.97M
DeepConvLSTM+SK convolution	93.15%	0.34M	10.42M

TABLE IX

INFERENCE SPEED ON GOOGLE NEXUS 6

General Models	Inference Time (ms/window)
Standard CNNs	144-165
SK networks (two-branch)	160-204
SK networks (three-branch)	162-208

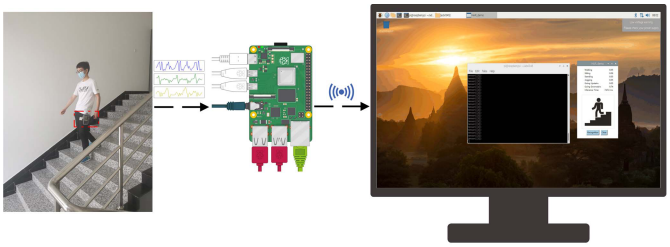


Fig. 11. Demo Application on Raspberry Pi 3B plus.

TABLE X

INFERENCE SPEED ON RASPBERRY PI 3B PLUS

General Models	Inference Time (ms/window)
Baseline	72.63-80.21
Baseline+SK convolution	105.23-130.42

SK block on the training set from WISDM and 2) load this trained model into the embedded platform, run it to read one sensor sample, and perform a real-time prediction. The timing is done after the model is loaded and starts to output a prediction. A Raspberry Pi-based program is developed for real-time activity recognition, and its UI is shown in Fig. 11. We set up Wi-Fi on the Raspberry Pi and wirelessly connect it to the remote computer. The Raspberry Pi platform with an accelerometer sensor ADXL345 is attached to the subject's front leg pocket. As can be seen from Table X, the baseline and the SK model take around 72.63–80.21 and 105.23–130.42 ms, respectively, to predict one window. In the case of WISDM, a 10-s window with a 95% overlapping rate is moved to segment sensor time series. Since the sliding step is equal to 500 ms, the recognition system needs to wait for 500 ms to process the next window. According to Table X, we could clearly observe that the proposed SK method can meet the runtime requirement on the embedded system.

V. CONCLUSION

In recent years, CNNs have become one dominant technique in the deep learning community, which results in appealing results in the HAR scenario. However, for most existing CNN architectures, the kernel size is usually fixed within the same feature layer, which fails to capture multiscale information from various human activities. In this article, a kernel selection approach is first proposed, which is able to aggregate multiscale information from multiple branches in the HAR scenario. A soft attention mechanism is utilized to adaptively fuse features from multiple kernels. Extensive experiments are conducted on several benchmark HAR datasets, which indicates that the SK convolution outperforms other SOTA methods with a similar budget in parameter and computation cost. In addition, we analyze the independent contribution of several crucial elements within kernel selection to better understand its mechanism.

ACKNOWLEDGMENT

The authors thank the graduate student Xing Wang for his help in the Raspberry PI experiment. They would like to express their sincere appreciation for the valuable suggestions provided by the editor and reviewers, which have considerably improved the quality of this article.

REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[2] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "From action to activity: Sensor-based activity recognition," *Neurocomputing*, vol. 181, pp. 108–115, Mar. 2016.

[3] T. Tuncer, F. Ertam, S. Dogan, and A. Subasi, "An automated daily sports activities and gender recognition method based on novel multikernel local diamond pattern using sensor signals," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9441–9448, Dec. 2020.

[4] Y. Zhang, G. Tian, S. Zhang, and C. Li, "A knowledge-based approach for multiagent collaboration in smart home: From activity recognition to guidance service," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 317–329, Feb. 2020.

[5] Z. Chen, S. Xiang, J. Ding, and X. Li, "Smartphone sensor-based human activity recognition using feature fusion and maximum full *a posteriori*," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 3992–4001, Jul. 2020.

[6] J. Suto and S. Oniga, "Efficiency investigation from shallow to deep neural network techniques in human activity recognition," *Cogn. Syst. Res.*, vol. 54, pp. 37–49, May 2019.

[7] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4278–4284.

[10] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 510–519.

[11] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[12] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. Esann*, vol. 3, 2013, p. 3.

[13] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, p. 1101, 2017.

[14] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 74–82, Dec. 2010.

[15] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.*, Jun. 2012, pp. 108–109.

[16] R. Chavarriaga *et al.*, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 2033–2042, Jan. 2009.

[17] M. Zeng *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput., Appl. Services*, 2014, pp. 197–205.

[18] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. IJCAI*, vol. 15, Buenos Aires, Argentina, 2015, pp. 3995–4001.

[19] K. Wang, J. He, and L. Zhang, "Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors," *IEEE Sensors J.*, vol. 19, no. 7, pp. 7598–7604, Sep. 2019.

[20] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," 2016, *arXiv:1604.08880*. [Online]. Available: http://arxiv.org/abs/1604.08880

[21] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1307–1310.

[22] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[23] C. Hu, Y. Chen, L. Hu, and X. Peng, "A novel random forests based class incremental learning method for activity recognition," *Pattern Recognit.*, vol. 78, pp. 277–290, Jun. 2018.

[24] H. Qian, S. J. Pan, B. Da, and C. Miao, "A novel distribution-embedded neural network for sensor-based activity recognition," in *Proc. IJCAI*, 2019, pp. 5614–5620.

[25] Q. Teng, K. Wang, L. Zhang, and J. He, "The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition," *IEEE Sensors J.*, vol. 20, no. 13, pp. 7265–7274, Jul. 2020.

[26] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 2, p. 679, 2018.

[27] Y. Tang, Q. Teng, L. Zhang, F. Min, and J. He, "Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors," *IEEE Sensors J.*, vol. 21, no. 1, pp. 581–592, Jan. 2021.

[28] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.

[29] J. Long, W. Sun, Z. Yang, and O. I. Raymond, "Asymmetric residual neural network for accurate human activity recognition," *Information*, vol. 10, no. 6, p. 203, Jun. 2019.

[30] I. Andrey, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, Jan. 2017.

[31] M. Zeng *et al.*, "Understanding and improving recurrent networks for human activity recognition by continuous attention," in *Proc. 2018 ACM Int. Symp. Wearable Comput.*, 2018, pp. 56–63.

[32] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 8–13.

[33] A. Khan, N. Hammerla, S. Mellor, and T. Plötz, "Optimising sampling rates for accelerometer-based human activity recognition," *Pattern Recognit. Lett.*, vol. 73, pp. 33–40, Apr. 2016.

[34] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, p. 106, 1962.

[35] J. I. Nelson and B. J. Frost, "Orientation-selective inhibition from beyond the classic visual receptive field," *Brain Res.*, vol. 139, no. 2, pp. 359–365, Jan. 1978.

[36] M. P. Sceniak, D. L. Ringach, M. J. Hawken, and R. Shapley, "Contrast's effect on spatial summation by macaque V1 neurons," *Nature Neurosci.*, vol. 2, no. 8, pp. 733–739, Aug. 1999.

[37] S. Aslani *et al.*, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, Aug. 2019.

[38] R. Kumar Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*. [Online]. Available: http://arxiv.org/abs/1505.00387

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

[40] Z. Wu *et al.*, "BlockDrop: Dynamic inference paths in residual networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8817–8826.

[41] X. Jia, B. De Brabandere, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," in *Proc. NIPS*, 2016, pp. 667–675.

[42] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[43] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," 2019, *arXiv:1901.10430*. [Online]. Available: http://arxiv.org/abs/1901.10430

[44] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[46] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[48] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7268–7277.

[49] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2018, pp. 7132–7141.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

Wenbin Gao received the B.S. degree from Changzhou Institute of Technology, Changzhou, China, in 2019. He is currently pursuing the M.S. degree with Nanjing Normal University, Nanjing, China.

His research interests include activity recognition, computer vision, and machine learning.

Lei Zhang received the B.Sc. degree in computer science from Zhengzhou University, Zhengzhou, China, in 2001, the M.S. degree in pattern recognition and intelligent system from Chinese Academy of Sciences, Beijing, China, in 2004, and the Ph.D. degree from Southeast University, Nanjing, China, in 2011.

He was a Research Fellow with the Institute for Pure and Applied Mathematics (IPAM), University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 2008. He is currently an Associate Professor with the School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing. His research interests include machine learning, human activity recognition, and computer vision.

Wenbo Huang received the B.S. degree from Nanjing University of Technology, Nanjing, China, in 2019. He is currently pursuing the M.S. degree with Nanjing Normal University, Nanjing.

His research interests include activity recognition, computer vision, and machine learning.

Fuhong Min received the master's degree from the School of Communication and Control Engineering, Jiangnan University, Wuxi, China, in 2003, and the Ph.D. degree from the School of Automation, Nanjing University of Science and Technology, Nanjing, China, in 2007.

From 2009 to 2010, she was a Post-Doctoral Fellow with the School of Mechanical Engineering, Southern Illinois University, Carbondale, IL, USA. She is currently a Professor with the School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing. Her research interests include circuits and signal processing.

Jun He received the Ph.D. degree from Southeast University, Nanjing, China, in 2009.

He was a Research Fellow with the Institute for Pure and Applied Mathematics (IPAM), University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 2008, and a Post-Doctoral Research Associate with The Chinese University of Hong Kong, Hong Kong, from 2010 to 2011. He is currently an Associate Professor with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing. His research interests include machine learning, computer vision, and optimization methods.

Aiguo Song (Senior Member, IEEE) received the Ph.D. degree in measurement and control from Southeast University, Nanjing, China, in 1996.

He is currently a Professor with the School of Instrument Science and Engineering, Southeast University. His current research interests include teleoperation, haptic display, the Internet Telerobotics, distributed measurement systems, and machine learning.

Dr. Song is also the Chair of the China Chapter of the IEEE Robotics and Automation Society.