

Plug-and-play multi-dimensional attention module for accurate Human Activity Recognition

Junjie Liang^{a,*}, Lei Zhang^{a,*}, Can Bu^a, Guangyu Yang^a, Hao Wu^b, Aiguo Song^c

^a School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, 210023, China

^b School of Information Science and Engineering, Yunnan University, Kunming, 650000, China

^c School of Instrument Science and Engineering, Southeast University, Nanjing, 210023, China

ARTICLE INFO

Keywords:

Human activity recognition
Convolutional neural networks
Multi-dimensional attention
Wearable sensors

ABSTRACT

Unlike image data, Human Activity Recognition (HAR) data is collected through sensors placed on different parts of the body. Each sensor may have varying importance for different human actions. Attention mechanisms can dynamically identify the significance of signals from different sensors. However, existing attention modules often operate along channel or spatial dimensions, producing one-dimensional or two-dimensional weights, and considering neurons within every channel or spatial location as equivalent. This might constrain their capacity to acquire more distinctive cues. In this paper, we present a novel Three-Dimensional Weight Attention Module (WAM) that considers both spatial and weight information of each channel. Specifically, we commence by introducing an energy function and subsequently assess the significance of each neuron through optimization. Unlike conventional channel or spatial attention modules, this module can compute 3D attention weights for feature maps without additional parameters. Experiments conducted on four diverse HAR datasets showcase that the Three-Dimensional Weight Attention Module can seamlessly integrate with convolutional models, significantly enhancing model performance without introducing extra computational burden. To further highlight the utility of WAM, we measure the actual inference time on an embedded platform.

1. Introduction

1.1. Background

Human activity recognition (HAR) based on wearable sensors [1] has emerged as a research hotspot over the past few years. At present, the ubiquitous utilization of inertial sensors (e.g., triaxial accelerometers and gyroscopes) for activity recognition has gained a lot of attention across various Internet-of-Things (IoT) scenarios like health-care, fitness, sports tracking, and smart homes [2–4]. On the other hand, convolutional neural networks (CNNs) [5–7] have achieved remarkable success in the field of computer vision. From the earliest LeNet [8] to the later VGGNet [9], ResNet [6], EfficientNet [10], GhostNet [11], these powerful CNN models have laid a solid basis for developing advanced HAR algorithms, which show a great potential in handling sensor data. Thanks to powerful feature extraction capabilities, deep CNNs have been widely adopted in a large variety of sensor-based HAR tasks [12,13], which has been increasingly integrated into various aspects of human daily life. Through alternately

stacking multiple operators such as convolution, pooling, activation, and other customized plug-and-play structures (termed as module in this paper), deep CNNs can automatically learn hierarchical features, which enable them to extract useful temporal and spatial information regarding intricate activities from raw sensor data, leading to precise recognition of complex human actions. In particular, as a beneficial complement to CNNs, attention mechanisms [14–17] show a great potential to enhance model performance by adaptively attending to various sensor modalities or time intervals, thereby providing a better interpretability for backbox deep model models in the context of HAR.

1.2. Limitation of prior art

At present, it still remains two important problems in HAR with CNN. First, the sensor signals inevitably contain some activity-irrelevant intervals. For example, the fall actions (Fig. 1) only happen in a relatively small time interval, instead of one entire window. Standard CNN alone usually fails to accurately detect the occurrence of fall

* Corresponding author.

E-mail addresses: JunjieLiang85@gmail.com (J. Liang), leizhang@njnu.edu.cn (L. Zhang), bucanya123@gmail.com (C. Bu), 211843010@njnu.edu.cn (G. Yang), haowu@ynu.edu.cn (H. Wu), a.g.song@seu.edu.cn (A. Song).

<https://doi.org/10.1016/j.comnet.2024.110338>

Received 18 October 2023; Received in revised form 19 January 2024; Accepted 12 March 2024

Available online 15 March 2024

1389-1286/© 2024 Elsevier B.V. All rights reserved.

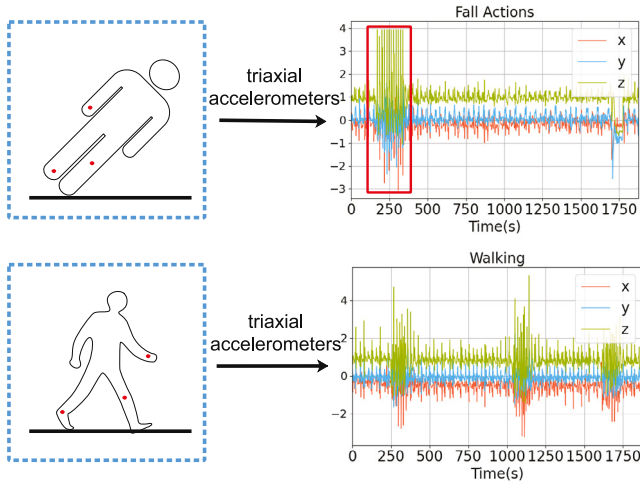


Fig. 1. Sensor signals under different body movements.

actions within such time series sensor signals. These activity-irrelevant intervals may potentially introduce disrupting noise while feeding the whole window into CNN. Second, different sensor modalities may be of different importance while recognizing various activities. For example, to detect activities during walking, the sensor modalities attached to the trunk is obviously less important than those on the foot or leg. It will also introduce substantial noise while applying unimportant sensors modalities in HAR. One main promising research line is to build attention modules [16–20], which can refine convolutional blocks while forcing the whole network to adaptively focus on more important features while attenuating activity-irrelevant information, yielding a more precise activity recognition. Such attention modules are often independent of the whole network, which could be easily integrated into a wide range of convolutional backbones. However, there are two main shortcomings for existing attention modules. First, previous most attention modules are only able to refine activity features along either channel or spatial (e.g., sensor modalities and time axes) dimensions alone, limiting the model’s adaptability and flexibility while inferring attention weights. Second, their designs heavily rely on a series of hand-crafted intervention factors (e.g., the selection of pooling), which usually requires additional parameters and consumes extra computational resources.

1.3. Research motivation

In this paper, to address the aforementioned issues, we propose a novel attention module, which is primarily built based on a well-established neuroscience theory. As illustrated in Fig. 2, the current mainstream attention modules usually focus on estimating either one-dimensional (Fig. 2(a)) or two-dimensional (Fig. 2(b)) attention weights to refine activity features. Different from prior most works, our attention module (Fig. 2(c)) exploits a fundamentally different idea through directly inferring three-dimensional attention weights, which can offer a more comprehensive importance selection mechanism to prioritize activity-relevant information and suppress activity-irrelevant disturbing signals for activity recognition. Such estimation of three-dimensional attention weights can enable the model to precisely capture more valuable cues that can be in well line with different activity labels. In stark contrast to traditional attention modules, our method is mainly inspired by the spatial suppression phenomenon of neuroscience observed in the mammalian brain, which can formulate the weight generation process as an energy function. Following previous neuroscience knowledge [17,21,22], we can easily derive a closed-form solution based on the energy function, which facilitates an efficient inference

for such three-dimensional attention weights. Thanks to the closed-form solution, our attention module is parameter-free, which requires no handcrafted design and contains no additional parameters. It can serve as a plug-and-play component to further boost the representational power of any convolutional backbones for activity recognition.

1.4. Main contribution

In summary, the proposed Three-Dimensional Weight Attention Module (WAM) is not only more flexible but also keeps a lightweight property for sensor-based HAR tasks. The main contributions of this paper are summarized as follows:

- In this paper, we introduce a plug-and-play three-dimensional weight attention module called WAM, which is parameter-free and can be seamlessly incorporated into any CNN architectures, enabling the model to automatically select important weights based on interior characteristics of multimodal sensor data. As a consequence, our attention module can better adapt to different HAR tasks.
- We evaluate the effectiveness and efficiency of the proposed WAM module on four widely employed HAR benchmark datasets including WISDM, PAMAP2, UniMiB-SHAR, and UCI-HAR. The WAM attention module has been shown to work well in the areas of HAR, which surpasses other state-of-the-art attention modules in terms of both accuracy and memory footprint. In addition, the practical inference time is tested on a real mobile device.
- The proposed WAM module is capable of directly estimating three-dimensional attention weights through designing an energy function. A closed-form solution can be derived based on the energy function, which significantly accelerates the weight calculation process. Such calculation of three-dimensional attention weights is more straightforward without requiring manual design, while maintaining a lightweight advantage.

Below, Section 2 first reviews previous research literatures and introduces different attention mechanisms. The main differences between them and our WAM approach are introduced. Section 3 then introduces the details of our approach. The main experimental results and a series of ablation studies are presented in Section 4. Finally, Section 5 summarizes this paper.

2. Related work

In this section, we primarily discuss previous research efforts, encompassing neural network architecture design and existing attention mechanisms.

2.1. Network architecture design

With its notable feature extraction capabilities and automatic learning abilities, deep CNNs have found widespread applications in the field of HAR based on wearable devices. The initial attempts can be traced back to scholars such as Zeng et al. [23], Ronao et al. [24], and Yang et al. [25], who employed CNN models to automatically extract discriminative features from time-series sensor data, enabling the precise identification of human activities. The VGGNet [9] proposed by Karen Simonyan et al. is one of the most widely applied neural network models currently. It is distinguished by its utilization of multiple small-sized convolutional kernels and deep-layer structures, enabling the attainment of more powerful feature representation. On the other hand, the ResNet model introduced by Kaiming He et al. [6] brought forward the concept of residual structures, addressing the issue of degradation in deep networks. This innovation allowed networks to become deeper and subsequently led to significant performance improvements across various tasks. Their research findings suggest that increasing network depth can significantly enhance the feature representation capabilities

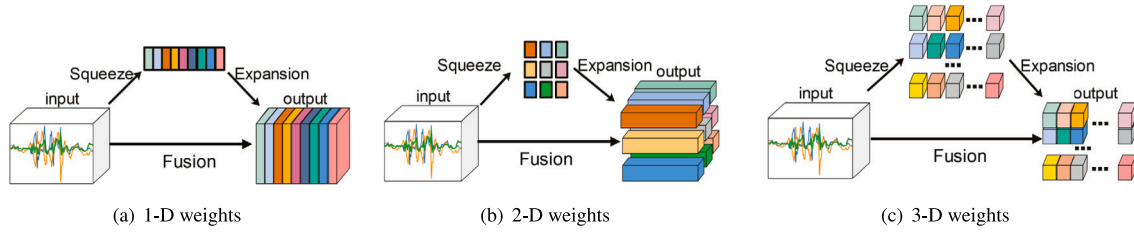


Fig. 2. Comparative analysis of various attention mechanisms reveals distinctions. Most existing attention modules primarily generate one-dimensional or two-dimensional weights from features, while our module directly estimates three-dimensional weights.

of CNN models. However, increasing the depth of the network also comes with high computational costs and memory resources. To overcome this challenge, many existing studies have focused on redesigning the convolutional structure to reduce the model's computational costs, such as employing strategies like depthwise separable convolution [26–28] and feature reuse [29,30]. In particular, the GhostNet [11] model effectively addresses the computational resource issue with its unique lightweight design. Nevertheless, this still may not be optimal. In contrast to the aforementioned works, our goal is to design a lightweight plug-and-play attention module that can be incorporated into different CNN models to further enhance the model's performance in HAR tasks.

2.2. Attention module design

Attention modules play a significant role in the field of HAR, aiding in enhancing a model's ability to recognize complex activities and extracting essential features from sensor data. Attention modules [17,18,31] enable models to focus more precisely on capturing crucial sensor signals and time steps, thereby improving the accuracy and efficiency of activity recognition. He et al. [32] innovatively introduced a hard attention mechanism by utilizing reinforcement learning methods to train activity recognition tasks, highlighting the most significant time intervals within long sensor sequences. In contrast, Wang et al. [33] introduced a soft attention mechanism that calculates weights to emphasize key time intervals while suppressing irrelevant or potentially interfering intervals. Similarly, Ma et al. [14] presented a framework that combines attention mechanisms with CNNs and gated recurrent unit networks to capture the spatial and temporal dependencies of sensor signals. There have also been other popular attention modules such as ECA [20], GC [19], and CBAM [18], and SE [16], which have dominated in deep learning community. Typically, SE proposed by Hu et al. first attempts to enhance feature representation by adaptively learning the weights for each channel from a global view. The attention module is further extended by other ideas, e.g., CBAM that adds spatial-wise attention. Recently, ECA [20] have empirically shown that it is important to avoid dimensionality reduction while learning the channel attention in SENet. Thus, they introduce a local cross-channel interaction strategy without performing dimensionality reduction, which can be efficiently implemented via adaptively selecting kernel size of one-dimensional convolution so as to determine coverage of local cross-channel interaction. On the other hand, sharing a similar idea to SENet, GC [19] have exploited a pioneering Non-Local Network (NLNet) to design a simplified network based on a query-independent formulation to capture long-range dependencies, which can preserve the same accuracy with NLNet but consume less computation. In summary, previous most attention algorithms have failed to distinguish different neurons within one channel or different neurons at one spatial position, thereby preventing the precise calculation of three-dimensional weights. Moreover, their attention modules are typically manually designed, which contain parameters and require substantial computational resources. In stark contrast, our attention module is designed based on well-established neuroscience theories [22,34,35], which are more interpretable in the context of HAR. It is worth noting that prior several attention modules have also been inspired by neuroscience, such as selective kernel

receptive field tuning and adaptive context-driven convolution [21]. Different from previous these works, our attention module is primarily motivated by an idea of spatial inhibition observed in the mammalian brain, which can formulate the attention weight generating process as an energy function. Since a fast closed-form solution can be easily derived based on the energy function, our attention module contains no parameters. Therefore, such compelling parameter-free property makes it clearly distinct from prior most attention works.

3. Propose method

Fig. 3 illustrates the application and implementation details of the proposed three-dimensional weight attention module in the context of HAR tasks. It is primarily divided into three sections, encompassing data acquisition, data preprocessing, and model construction. This section provides an overview of the construction approach of the introduced attention module.

As mentioned earlier, the direct inference of three-dimensional weights is a highly challenging task. In order to implement three-dimensional attention [17,36,37], it becomes necessary to individually estimate the significance of each neuron. Specifically, neurons with pronounced spatial inhibitory features should be assigned higher priority. The straightforward approach to pinpoint high-priority neurons involves gauging the linear separability between the target neuron and its counterparts. Consequently, we establish the following energy function for every neuron:

$$e_n(w_n, b_n, \mathbf{y}, x_i) = (y_n - \hat{n})^2 + \frac{1}{S-1} \sum_{i=1}^{S-1} (y_o - \hat{x}_i)^2 \quad (1)$$

$$\hat{n} = w_n n + b_n, \quad \hat{x}_i = w_n x_i + b_n,$$

where n and x_i are the target neuron and other neurons in a single channel of the input feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$. i represents the index along the spatial dimensions, while $S = H \times W$ indicates the number of neurons within the current channel. w_n and b_n correspond to the weighted sum and bias of the transformation. When $\hat{n}$ is equal to y_n , $(y_n - \hat{n})^2 = 0$, the Eq. (1) reaches its minimum value. By minimizing Eq. (1), we can establish the linear separability between the target neuron n and the remaining neurons within the same channel. For computational convenience, we treat y_n and y_o as binary labels (i.e., -1 and 1) and incorporate a regularization term into Eq. (1). Consequently, the final energy function can be represented as follows:

$$e_n(w_n, b_n, \mathbf{y}, x_i) = \frac{1}{S-1} \sum_{i=1}^{S-1} (-1 - (w_n x_i + b_n))^2 + (1 - (w_n n + b_n))^2 + \lambda w_n^2 \quad (2)$$

In fact, there are S neurons on each channel, with each neuron corresponding to an energy function. Therefore, S energy functions need to be computed for each channel, which is resource-intensive. Fortunately, we can easily solve the closed-form solution for Eq. (2) with respect to

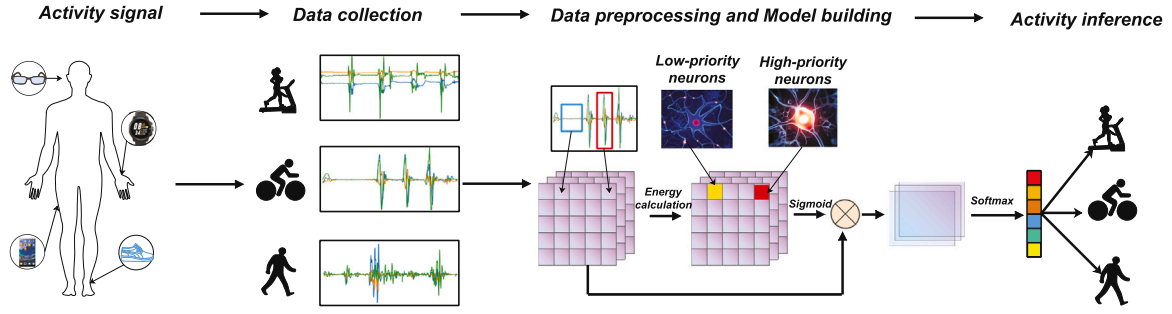


Fig. 3. Network diagram of WAM. The curve represents the sensor time series.

w_n and b_n using the following approach:

$$w_n = \frac{2(n - \mu_n)}{(n - \mu_n)^2 + 2\sigma_n^2 + 2\lambda}, \quad (3)$$

$$\mu_n = \frac{1}{S-1} \sum_{i=1}^{S-1} x_i, \quad \sigma_n^2 = \frac{1}{S-1} \sum_{i=1}^{S-1} (x_i - \mu_n)^2, \quad (4)$$

where σ_n^2 and μ_n respectively represent the variance and mean calculated over all neurons except the current neuron n . Considering that the solutions in Eqs. (3) and (4) are obtained within an individual channel, we can assume that all pixels within an individual channel adhere to the same distribution. Building upon this assumption, calculating the variance and mean for all neurons and subsequently reusing them across all neurons within the channel becomes viable [17,21,38]. This approach not only substantially reduces computational costs but also eliminates the need for redundant calculations of μ and σ for each neuron. Therefore, the minimum energy can be determined using the following formula:

$$e_n^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(n - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda}, \quad (5)$$

$$\hat{\mu} = \frac{1}{S-1} \sum_{i=1}^{S-1} x_i, \quad \hat{\sigma}^2 = \frac{1}{S-1} \sum_{i=1}^{S-1} (x_i - \mu_n)^2.$$

Eq. (5) reveals that as the energy e_n^* decreases, the similarity between neuron n and its neighboring neurons decreases. In other words, the features represented by neuron n become more distinctive, which is crucial for extracting prominent features. As a result, the significance of each neuron can be represented by evaluating the magnitude of $1/e_n^*$. Our approach aligns with the investigation of semantic part matching as conducted by [39]. Nevertheless, their approach involves computing large covariance matrices, which is unsuitable for deep neural networks. In contrast to [39], our methodology centers around the individual neuron and incorporates this linear separability into a complete end-to-end framework. Additionally, we offer extensive insights derived from the field of neuroscience.

Until now, we have proceeded by deriving an energy function and consequently establishing the prioritization (significance) of each neuron. In line with the research conducted by [17,21,22,40], the modulation of attention within the mammalian brain often involves a scaling effect, commonly referred to as gain, which impacts neuronal responses. As a result, during the process of refining features, we opt for the utilization of a scaling operator instead of addition. Our module's ultimate expression can be succinctly formulated as follows:

$$\tilde{\mathbf{X}} = \text{sigmoid}\left(\frac{1}{E}\right) \odot \mathbf{X}, \quad (6)$$

where E represents the aggregation of e_n^* across spatial and channel dimensions. The addition of the sigmoid function aims to constrain excessively large values within E . Due to the monotonic nature of the

Algorithm 1: Algorithm of WAM

Input: Input feature $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$, Coefficient λ in

Eq. (5).

Output: Attended feature tensor $\tilde{\mathbf{X}}$.

- 1: Set: $t = 0$, $z = H \times W - 1$, $m = \Phi$.
- 2: **for** $i = 1 \rightarrow H \times W - 1$ **do**
- 3: Calculating the variance of a neuron:
- 4: $v = (\mathbf{X} - H_i \times W_i)^2$,
- 5: Add v into m ,
- 6: $t = t + v$,
- 7: **end for**
- 8: Calculate the variance of the channel: $\frac{t}{z}$.
- 9: Calculate the importance of all neurons by Eq. (5):
- 10: $e = \frac{m}{(4 * (t/z + \lambda))} + 0.5$.
- 11: **return** $\tilde{\mathbf{X}} = \text{sigmoid}(e) \odot \mathbf{X}$.

sigmoid function, it does not affect the relative importance of each neuron.

To summarize, our three-dimensional attention module algorithm, as depicted in Algorithm 1. Through iterative optimization, we minimize the energy function to find the optimal linear classification function formed by w_n and b_n . A smaller value of this function indicates a higher degree of linear separability between the current neuron n and the rest of the neurons x_i . In other words, the current neuron possesses a higher priority. Thus, we can use $1/e_n^*$ to represent the importance of neurons.

4. Experiments

In this section, we conducted a performance evaluation of the 3D weight attention module on four different HAR datasets. Compared to other types of attention modules, our module exhibited superior performance, effectively capturing a diverse range of human activity information with greater accuracy. By comparing with classical attention modules and benchmark models, we validated the exceptional performance of the 3D weight attention module. Across various performance metrics, our 3D weight attention module demonstrated significant enhancements, further affirming its effectiveness in HAR tasks. Through visualization analysis using heatmaps, we showcased the activation patterns of the 3D weight attention module on features. Finally, by measuring real-world activity execution time on the Raspberry Pi 4 B+ platform, we demonstrated that our module not only improves model performance but also incurs no additional computational overhead.

4.1. Dataset description

We conducted experiments on four common HAR datasets, namely WISDM [41], PAMAP2 [42], UniMiB-SHAR [43], and UCI-HAR [44].

Table 1
Data preprocessing (A=accelerometer, G=gyroscope, M=magnetometer).

Dataset	Attribute	Activity classes	Windows size	Overlap rate	Sampling rates	Sensor types	Number of subjects
WISDM		6	200	95%	20 Hz	A	29
PAMAP2		12	171	50%	100 Hz	A, G, M	9
UniMiB-SHAR		17	151	–	50 Hz	A	30
UCI-HAR		6	128	50%	50 Hz	A, G	30

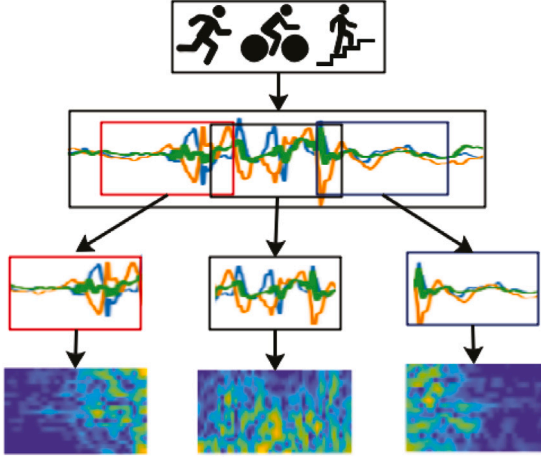


Fig. 4. Processing of sensor time series data through sliding window technology.

These datasets consist of data collected from multiple sensors positioned at various parts of the body in different environments and conditions. Detailed information about the data preprocessing process, including aspects such as sampling rates and sensor categories, is provided in Table 1. Among the existing HAR data preprocessing methods, the sliding window technique has been widely adopted due to its simplicity and efficiency. As shown in Fig. 4, the sensor time series are segmented using fixed-length sliding windows, with overlapping allowed between adjacent windows to maintain continuity in the sensor signals. Through the sliding window technique, we effectively mitigate the influence of noise signals, normalize the data, and segment it into appropriate windows for subsequent feature extraction and analysis.

WISDM dataset: This dataset was collected by the WISDM research team at Fordham University with the assistance of 29 participants. Each participant was instructed to perform six activities, including ‘Walking’, ‘Going Downstairs’, ‘Going Upstairs’, ‘Standing’, ‘Running’ and ‘Sitting’. The raw sensor data was obtained from the three-axis accelerometer in Android smartphones. During the data collection process, participants carried the Android smartphone in their pockets. The accelerometers captured activity data at a sampling rate of 20 samples per second, with measurements taken every 50 ms. The entire dataset was randomly divided into three parts, with 70% used for model training, 10% for validation, and 20% for testing the model’s performance [41].

PAMAP2 dataset: This dataset was collected and released by the research team at Ruhr-Bochum University. The PAMAP2 dataset includes sensor data from 9 participants and covers a variety of daily activities. The sensors included in the dataset are three-axis accelerometers, three-axis gyroscopes, and magnetometer sensors, which record participants’ motion and physiological signals during different activities. The dataset contains activity categories such as ‘Going Upstairs’, ‘Going Downstairs’, ‘Walking’, ‘Running’, ‘Standing’ and ‘Cycling’ [42].

UniMiB-SHAR dataset: This dataset contains 11,771 activity samples collected from 30 participants. The sensor data was sampled at a fixed frequency of 50 Hz using embedded three-axis BMA220 accelerometers worn in the front pocket of participants. All activity samples are divided into 17 categories, including ‘Standing’, ‘Walking’, ‘Clapping’, ‘Falling Forward’, ‘Falling Backward’ and others [43].

Table 2
Implementation details.

	WISDM	PAMAP2	UniMiB-SHAR	UCI-HAR
Kernel size	(7×1)	(9×1)	(5×1)	(9×1)
Stride	(1×1)	(4×1)	(1×1)	(2×1)
Padding	(3×0)	(4×0)	(2×0)	(4×0)
BN	False	True	False	True
Learning rate	0.001	0.001	0.001	0.001
Optimizer	SGD+0.9	Adam	SGD+0.9	Adam
Batch size	64	256	64	256

UCI-HAR dataset: This dataset was collected and released by researchers at the University of California, Irvine (UCI). The UCI-HAR dataset includes sensor data from 30 participants, covering six different activities from daily life, including ‘Walking’, ‘Lying Down’, ‘Going Upstairs’, ‘Going Downstairs’, ‘Sitting’ and ‘Standing’. The sensors in the dataset include accelerometers and gyroscopes, which record participants’ motion signals during various activities. Each sample in the dataset consists of sliding windows with a time step of 128, where each time step contains 9 features. These features include readings from three-axis accelerometers and three-axis gyroscopes. Each sample is labeled with an activity category to indicate the corresponding activity [44].

4.2. Implementation details

We conducted extensive experiments on four distinct HAR datasets. To accommodate the specific characteristics of each dataset, we flexibly adjusted the parameter settings. In terms of the experiment details, Table 2 presents comprehensive information. For the WISDM and UniMiB-SHAR datasets, during the training phase, we utilized the SGD optimizer with a momentum of 0.9. The initial learning rate was set to 0.001, and it was reduced by a factor of 10 every 50 epochs. Simultaneously, the batch size was set to 64 to ensure efficient parallel computation. On the other hand, when working with the UCI-HAR and PAMAP2 datasets, we utilized the Adam optimizer with a batch size of 256 during the training of the model. Similarly, the initial learning rate was set to 0.001, and it was reduced by a factor of 10 every 50 epochs to optimize the model training process.

Unlike the SE module [16], our 3D weight attention module measures the priority of neuron weights through an energy function. This parameter-free process enables our attention module to update existing network architectures without increasing computational costs. Extensive experiments were conducted to evaluate the performance gains brought by the 3D weight attention module. To explore the generalization ability of the proposed attention module, we utilize a standard three-layer CNN as our backbone network. Fig. 5 illustrates an overall architecture of the three-layer CNN, which contains 64, 128, and 256 filters/kernels respectively. Both the batch normalization (BN) operation and Rectified Linear Unit (ReLU) activation function are inserted after each convolution block and then that output is max pooled (Maxpooling). Since there have been heterogeneous sensor modalities, sampling rates, and sliding windows during data collection, we apply various kernel sizes and sliding strides across the four different HAR datasets (also see Table 2 for more details). In fact, the proposed attention module can be inserted between any two convolutional layers. Without loss of generality, we insert the attention module after the

Table 3
Accuracy & Parameters & FLOPs of models on various datasets (see Refs. [14,45–51]).

Method \ Dataset	WISDM	PAMAP2	UniMiB-SHAR	UCI-HAR
Baseline	97.13%&0.30M&17.14M	91.56%&0.43M&69.67M	74.62%&0.23M&9.14M	96.84%&0.40M&21.85M
+ SE	97.22%&0.32M&17.17M	92.15%&0.45M&69.72M	74.84%&0.25M&9.16M	96.89%&0.41M&21.87M
+ WAM(ours)	97.36%&0.30M&17.14M	92.36%&0.43M&69.67M	76.18%&0.23M&9.14M	97.51%&0.40M&21.85M
Other Result	93.32% [45] 97.50% [49]	91.16% [46] 89.30% [14]	76.14% [47] 75.67% [50]	97.10% [48] 96.91% [51]

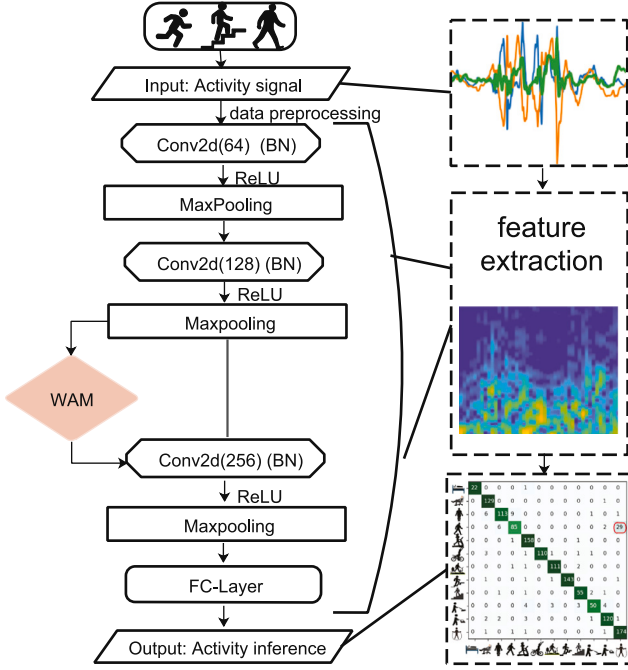


Fig. 5. Brief description for architecture.

second convolutional block. In particular, our module is simple to implement along with other existing backbone networks. To showcase the flexibility of the proposed 3D Weight Attention Module, we also integrate it into stronger residual networks and existing lightweight GhostNet backbones, which will be detailed later.

4.3. Main results

Most existing channel attention modules utilize average pooling or max pooling to capture global information within channels. While these pooling operations help models learn the overall importance of each channel, they may have limitations when dealing with fine-grained details and correlations within channels. However, our weight attention module offers a potential solution to this issue. In contrast to traditional pooling operations, WAM employs an attention fusion strategy. Specifically, the final attention weight is generated by the weighted combination of channel and spatial attention weights. This fusion strategy enables the network to flexibly adjust attention distribution across different dimensions.

In this section, we evaluate the effectiveness of the proposed 3D Weight Attention Module on four benchmark HAR datasets. As shown in Table 3, we compare WAM with traditional channel attention modules based on evaluation metrics including accuracy, parameter count, and FLOPs. Visualizations of the testing accuracy curves for the baseline CNN model, SE attention module [16], and WAM attention module are presented in Fig. 6, intuitively demonstrating the performance advantage of our attention module. Experimental results indicate that the 3D Weight Attention Module significantly enhances the activity recognition capability of the original baseline CNN while maintaining

Table 4
WAM performance in different network models.

Model	ResNet		Ghost	
	Baseline	+WAM	Baseline	+WAM
WISDM	97.33%	97.64%	97.01%	97.29%
PAMAP2	92.17%	93.03%	91.54%	92.12%
UniMiB-SHAR	76.73%	77.91%	74.11%	75.42%
UCI-HAR	97.01%	97.72%	96.34%	96.81%

the same FLOPs and parameter count. By introducing the 3D Weight Attention Module, we effectively enhance the model's feature selection ability, allowing it to focus more on critical neurons and features. Taking the WISDM dataset as an example, our 3D Weight Attention Module achieved a 0.23% increase in accuracy without introducing any additional computational cost. More notably, when dealing with the PAMAP2 and UniMiB-SHAR datasets, our module demonstrated performance improvements of 0.80% and 1.56%, respectively. Meanwhile, compared to traditional attention modules like the SE Module, our module exhibited significant advantages in terms of performance enhancement. For instance, on the UCI-HAR dataset, the traditional channel attention module only brought a marginal 0.05% improvement in performance, while our 3D Weight Attention Module achieved a remarkable 0.67% enhancement. These results strongly demonstrate the effectiveness and superiority of the proposed 3D Weight Attention Module, providing robust support for performance improvements in the field of HAR.

4.4. Exploring the flexibility of WAM

To validate the versatility of our proposed approach, we seamlessly incorporated the attention module into two widely utilized neural network architectures: ResNet [6] and Ghost models [11]. For the ResNet model, we directly incorporated the Weighted Attention Module after each residual block, enabling the enhancement of model performance through attention mechanisms without altering the original network structure. Additionally, we replaced the convolutional blocks in a three-layer CNN model with Ghost Modules, followed by the introduction of the WAM module after each Ghost Module. Through this adaptable integration approach, we validated the deployability of WAM across distinct architectures, further confirming its versatility and performance stability. The experimental outcomes, as illustrated in Table 4, demonstrate the effectiveness of our method across four diverse HAR datasets. Notably, for various models and HAR datasets, our attention module yielded significant performance improvements. For instance, on the UniMiB-SHAR dataset, our module achieved performance gains of 1.18% and 1.31% for the ResNet and Ghost models, respectively.

4.5. Feature map visualization

Visualizing three activity samples including 'Sitting Down', 'Running', and 'Jumping' on the UniMiB-SHAR dataset further demonstrates the effectiveness of the proposed approach. As shown in Fig. 7, we visualize the feature maps generated by both the SE module and the 3D attention module. In HAR tasks, visualizing feature maps is crucial for understanding how the model works and the effectiveness of the attention mechanism. By observing the feature maps, we can intuitively

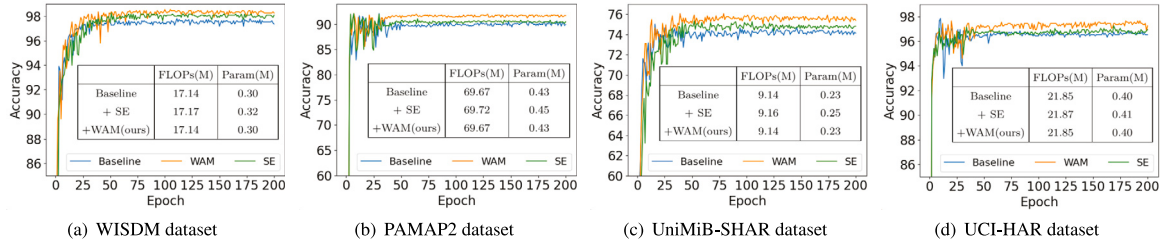


Fig. 6. Test accuracy curves with both backbone architectures at four HAR benchmarks.

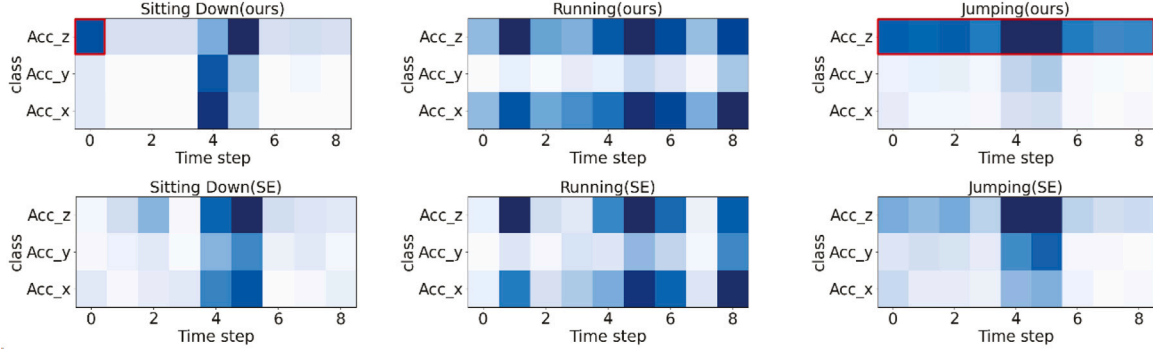


Fig. 7. Visualizing the 3-D attention on three activity samples, i.e., 'Sitting down', 'Running' and 'Jumping'.

comprehend the areas that the model focuses on in different activity samples, enabling a deeper analysis of the impact of attention mechanisms on model performance. In the figure, Acc_x , Acc_y , and Acc_z correspondingly indicate the acceleration signals in the forward, lateral, and upward directions. This visualization provides insights into how the attention modules guide the model's attention to specific regions, enhancing the model's ability to capture activity-related features.

Firstly, we considered the activity 'Sitting Down'. In comparison to activities like Running and Jumping, this is a relatively static activity. In Fig. 7, we present the feature responses of both the SE module and the 3-D attention module on this activity sample. It can be observed that while the SE module captures the important information within the activity, it overlooks many intricate details. For instance, during the moment of sitting down, the sensor signals on the z -axis are particularly sensitive, yet the SE module fails to attend to this crucial information. In contrast, our 3-D attention module not only retains significant information but also amplifies detailed features.

Lastly, we investigated the complex dynamic activities of 'Running' and 'Jumping'. Taking 'Jumping' as an example, it involves rapid limb movements, necessitating the model to capture swift changes in motion. 'Jumping' primarily entails vertical movement along the z -axis, making the z -axis sensor signal most significant. As seen in Fig. 7, our attention module concentrates more on the z -axis sensor signal. This further substantiates the effectiveness of our proposed 3-D attention module in capturing high-speed and momentary action features. Through visual analysis, we gain a more comprehensive understanding of the significance and superiority of the 3-D attention module in the field of HAR.

4.6. t-SNE visualization of clustering results

We employed the features of the post-trained attention model for t-SNE [52] clustering. Specifically, we utilized the trained attention model to extract feature vectors from data points, which were then input into the t-SNE algorithm for dimensionality reduction and clustering. Through this approach, we could observe the distribution and clustering relationships of the features learned by the attention model in the low-dimensional space, as well as the similarities and distinctions

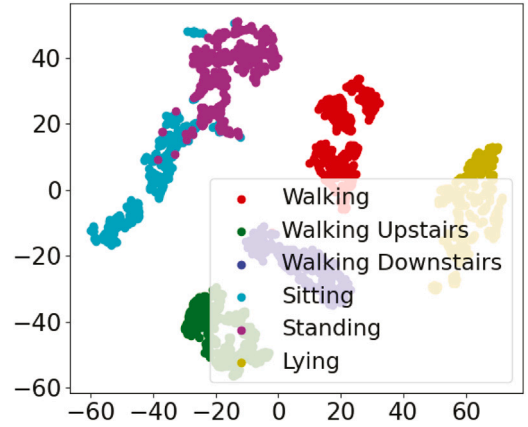


Fig. 8. Observing data similarity through t-SNE visualization.

among different action categories. We conducted t-SNE clustering on the UCI-HAR dataset, and the results are depicted in Fig. 8. Similar to the confusion matrix in Fig. 9, due to both 'Standing' and 'Sitting' belonging to the static action category, the highest similarity was observed between 'Standing' and 'Sitting' in the figure. Through t-SNE clustering, we not only discerned the similarity among static actions but also facilitated the differentiation between static and dynamic actions. This aids the model in better comprehending various facets of the data, thereby guiding the model to identify different human activities more accurately.

4.7. Confusion matrix

The confusion matrix, as a crucial tool for evaluating classification models, offers a comprehensive analysis of the classification results across different categories for various models. By visualizing the confusion matrix, we can intuitively grasp the predictive outcomes of different models on each category. Within this section, we conducted an assessment of the effectiveness of both our attention module and

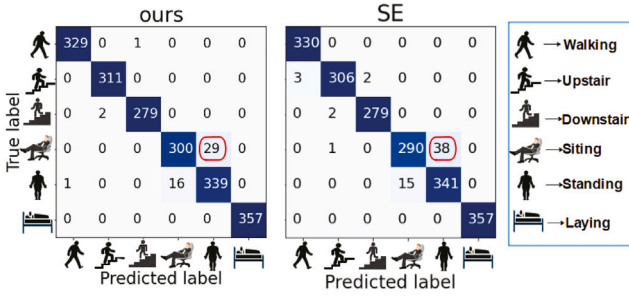


Fig. 9. Confusion matrices on UCI-HAR dataset.

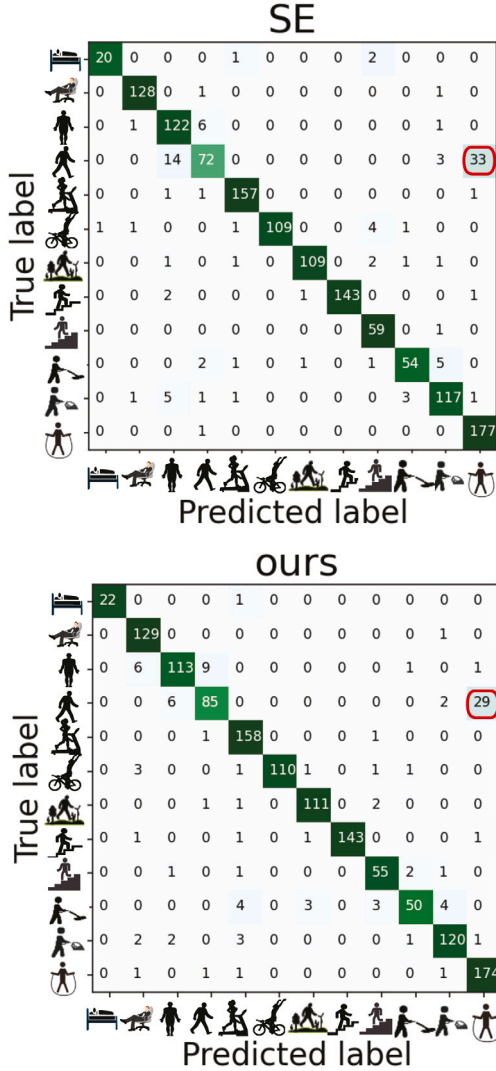


Fig. 10. Confusion matrices on PAMAP2 dataset.

SE module across diverse action categories within the UCI-HAR and PAMAP2 datasets.

We first evaluated the classification performance of the SE module and the WAM module on the UCI-HAR dataset. The experimental results in Fig. 9 demonstrate that actions with similarities are prone to be incorrectly predicted by the model. For instance, the static actions of ‘Standing’ and ‘Sitting’ share many similar sensor signal features, leading to confusion between these two activities. In the predictions of the SE attention model, 38 activity samples of ‘Sitting’ were mistakenly classified as standing. However, in our attention model, only 29 activity

samples were misclassified. This visually illustrates the superiority of our attention module in HAR.

Lastly, we conducted an evaluation of the classification performance of various action categories in the PAMAP2 dataset under different attention mechanisms. The experimental results, as shown in Fig. 10, intuitively illustrate the predictive outcomes of 12 distinct action categories across different attention models. The main diagonal indicates the number of correctly predicted samples for each category. The results clearly indicate that the SE attention model misclassified 33 ‘walking’ action samples as ‘Skipping Rope’. This misclassification can be attributed to the similarity in signal waveforms between ‘Walking’ and ‘Skipping Rope’, both being dynamic action categories. However, our attention module excels not only in distinguishing static action features in the UCI-HAR dataset but also demonstrates its effectiveness in recognizing dynamic action features in the PAMAP2 dataset. In comparison to the SE attention model, our three-dimensional attention model misclassified only 29 activity samples.

4.8. Compared to the spatial attention module

The CBAM [18] attention module combines channel attention and spatial attention, primarily consisting of two sub-modules: the channel attention module and the spatial attention module. Within the CBAM attention module, channel attention primarily serves to recalibrate the feature map weights across distinct channels, whereas spatial attention is harnessed to recalibrate feature map weights across diverse spatial positions. Although the CBAM attention mechanism has achieved significant success in terms of classification performance, it has not fully considered the information exchange between different channels. In contrast, our WAM emphasizes fine-grained control over channel information and inter-channel information exchange through the adoption of a selective and interactive multi-dimensional attention mechanism. This approach allows the WAM to retain critical features while also promoting feature diversity and richness, thereby further enhancing the convolutional model’s performance in HAR. In this section, we compared the classification performance of different attention modules on four different HAR datasets, further substantiating the superior performance of our attention module in the field of HAR. It is noteworthy that, due to our attention module not introducing any additional parameters, we compared the CBAM attention model and our model with similar parameter settings. From the experimental results shown in Table 5, it is evident that our attention module outperforms the CBAM attention module under similar parameter configurations. For instance, on the UniMiB-SHAR dataset, our module outperforms the CBAM attention module by 0.95% in terms of classification performance (76.18% vs. 75.23%). This further validates the superiority of our proposed attention module.

In addition to CBAM, we also compare the proposed method with two other representative attention modules in Table 5. Generally speaking, all these attention modules could consistently improve the baseline model by a clear margin. For example, comparing to ECA [20] and GC [19], the proposed WAM module can obtain superior results of 97.36%, 92.36%, 76.18%, and 97.51% on those datasets including WISDM, PAMAP2, UniMiB-SHAR, and UCI-HAR, respectively. As aforementioned, both ECA and GC always treat either all neurons at one spatial position or all neurons within one channel equally, so that they could not effectively calculate true three-dimensional attention weights. Moreover, their algorithms have to rely on laborious or time-consuming manual design such as the choice for pooling. As summarized in Table 6 [16–20], it can be seen that the choice of these structure designs requires significant engineering. In contrast, the proposed WAM module is primarily built on a well-known neuroscience theory (Eq. (5)) without involving human intervention, which can help the network to learn more discriminative neurons by directly calculating true three-dimensional attention weights, which is more

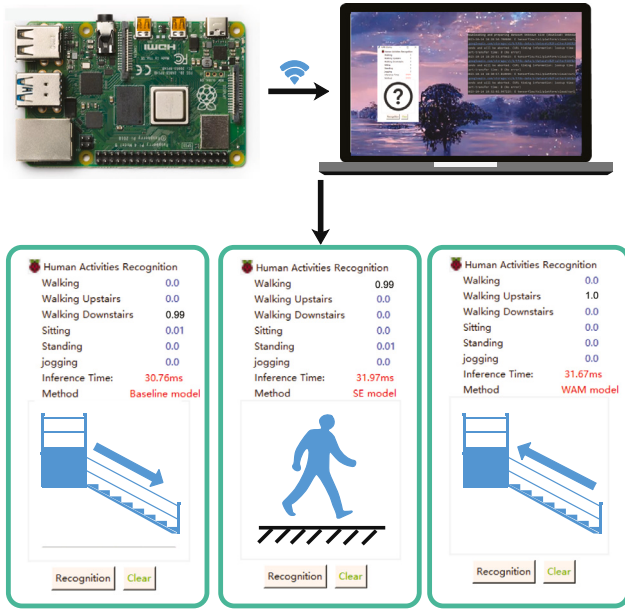


Fig. 11. Real-time HAR system on Raspberry Pi 4 B+.

Table 5

The performance comparisons of different attention modules.

Method	Accuracy	FLOPs	#Params	+ Params(M)
WISDM				
Baseline	97.13%	17.14	0.30	0
+ECA	97.15%	17.15	0.31	0.01
+GC	97.20%	17.18	0.33	0.03
+CBAM	97.22%	17.17	0.32	0.02
+WAM(ours)	97.36%	17.14	0.30	0
PAMAP2				
Baseline	91.56%	69.67	0.43	0
+ECA	91.64%	69.71	0.43	0
+GC	91.77%	69.72	0.44	0.01
+CBAM	91.81%	69.72	0.45	0.02
+WAM(ours)	92.36%	69.67	0.43	0
UniMiB-SHAR				
Baseline	74.62%	9.14	0.23	0
+ECA	74.67%	9.16	0.24	0.01
+GC	74.95%	9.16	0.24	0.01
+CBAM	75.23%	9.16	0.24	0.01
+WAM(ours)	76.18%	9.14	0.23	0
UCI-HAR				
Baseline	96.84%	21.85	0.40	0
+ECA	96.89%	21.86	0.40	0
+GC	96.99%	21.87	0.41	0.01
+CBAM	97.06%	21.87	0.41	0.01
+WAM(ours)	97.51%	21.85	0.40	0

interpretable. In particular, thanks to its closed-form solution, the proposed WAM module is parameter-free, which can effectively strengthen the representation power in various HAR tasks without incurring any additional parameters, thereby taking an obvious advantage over other attention modules.

v v v

4.9. Runtime analysis

For practical considerations, we deploy the proposed method to measure the actual running time on an embedded platform with limited resources. In order to realize a real-time HAR system on an embedded platform, we implement it through the following two steps: firstly,

Table 6

Comparisons of different attention modules in structural design. Several involved operators are as follows: Spatial Average-Pooling (SpatialAP), Spatial Max-Pooling (SpatialMP), Channel Average-Pooling (ChannelAP) or Channel Max-Pooling (ChannelMP), Normal Convolution (2-DC), Channel-based Convolution (1-DC), Fully-Connected Layers (FC), Batch Normalization (BN), Layer Normalization (LN), Softmax and Rectified Linear Unit (ReLU).

Method	Operators	Structural design
ECA	SpatialAP, 1-DC	handcrafted
GC	1-DC, LN, Softmax, FC Layer, ReLU	handcrafted
CBAM	SpatialAP, SpatialMP, ReLU, FC Layer, ChannelAP, ChannelMP, BN, 2-DC	handcrafted
ours	SpatialAP, /, $\odot$, +	Eq. (5)

Table 7

Inference time statistic.

Dataset	Method	Inference time (ms)
WISDM	Baseline	30.76
	SE model	31.97
	WAM model	31.67

we train a deep convolutional model on the WISDM dataset using the attention module; then we install the trained model into the embedded platform and run the deep model in the embedded platform to obtain the actual inference time of each active sample. We chose the Raspberry Pi 4 B+ as our test platform and developed a python-based activity recognition application with the user interface shown in Fig. 11, which contains graphical representations and predicted probabilities for all the activity categories of the WISDM dataset. The activity classification program loads the best weights and biases into the test platform and evaluates the prediction accuracy. Building one 6-class classifier, I run a Softmax function to identify which activity class has been predicted with the highest confidence. From left to right, the data in Fig. 11 illustrates three compared neural networks (i.e., the standard CNN baseline, SE attention network, and our WAE model) accepting inputs of different activity samples (i.e., ‘Walking Downstairs’, ‘Walking’, and ‘Walking Upstairs’) and generating their predicted outputs of (0.0, 0.0, 0.99, 0.01, 0.0, 0.0), (0.99, 0.0, 0.0, 0.0, 0.01, 0.0), and (0.0, 1.0, 0.0, 0.0, 0.0, 0.0) respectively, where the highest confidence score is highlighted by black font. In addition, we measure inference time of PyTorch models using “time” library, where a timer is launched (i.e., ‘starting_time’) when the pre-trained model starts to read new data, and then be terminated once a predictive output is returned (i.e., ‘ending_time’). A simple subtraction equation is used to calculate the time elapsed during inference (i.e., 30.76 ms, 31.97 ms, and 31.67 ms). Fig. 12 illustrates the inference times for different attention modules in a sample of 200 activities, and Table 7 shows the actual inference times for different attention modules in a random sample. The experimental results show that our attention modules improve the performance while adding essentially no additional computational overhead compared to the baseline model (30.76 ms vs. 31.67 ms). This further demonstrates the utility and performance superiority of our attention module on embedded platforms.

5. Conclusion

Drawing inspiration from neuroscience theories, this paper introduces an innovative WAM attention module. Building upon existing spatial inhibition theories, we formulate an energy function and optimize it to unveil the significance of each neuron. To accelerate weight computations, we employ a closed-form solution to solve the energy function. It is important to emphasize that our attention module does

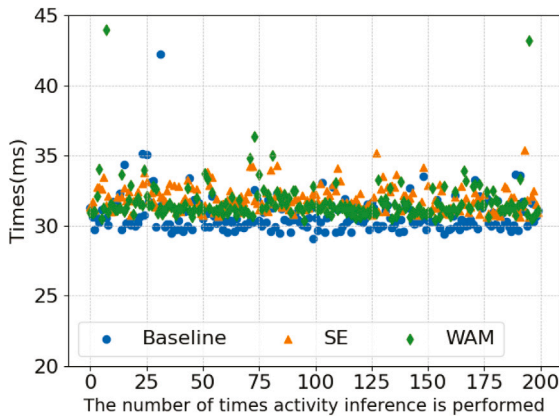


Fig. 12. Inference speed over 200 times.

not introduce any additional parameters, making it highly advantageous for practical HAR applications. Extensive experimentation validates the effectiveness and versatility of the proposed attention module. Across various HAR datasets, our attention module consistently outperforms other attention mechanisms under the same parameter settings. Lastly, through experimentation on embedded devices, we further confirm the feasibility of applying this attention module in real-world scenarios.

CRedit authorship contribution statement

Junjie Liang: Methodology, Investigation, Data curation, Conceptualization. **Lei Zhang:** Writing – review & editing, Writing – original draft. **Can Bu:** Software, Formal analysis. **Guangyu Yang:** Validation, Formal analysis. **Hao Wu:** Supervision, Methodology. **Aiguo Song:** Supervision, Investigation.

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.

Data availability

No data was used for the research described in the article.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China under Grant 62373194 and in part by the Natural Science Foundation of Jiangsu Province, China under grant BK20191371. Lei Zhang is the corresponding author.

References

- [1] Y. Bai, L. Lu, J. Cheng, J. Liu, Y. Chen, J. Yu, Acoustic-based sensing and applications: A survey, *Comput. Netw.* 181 (2020) 107447.
- [2] A. Farrokhi, R. Farahbakhsh, J. Rezazadeh, R. Minerva, Application of internet of things and artificial intelligence for smart fitness: A survey, *Comput. Netw.* 189 (2021) 107859.
- [3] F. Xiao, Q. Miao, X. Xie, L. Sun, R. Wang, SHMO: A seniors health monitoring system based on energy-free sensing, *Comput. Netw.* 132 (2018) 108–117.
- [4] H. Ning, F. Shi, T. Zhu, Q. Li, L. Chen, A novel ontology consistent with acknowledged standards in smart homes, *Comput. Netw.* 148 (2019) 101–107.
- [5] S. Liu, W. Deng, Very deep convolutional neural network based image classification using small training sample size, in: 2015 3rd IAPR Asian Conference on Pattern Recognition, ACPR, 2015, pp. 730–734.

- [6] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 630–645.
- [7] S.-M. Lee, S.M. Yoon, H. Cho, Human activity recognition from accelerometer data using convolutional neural network, in: *2017 IEEE International Conference on Big Data and Smart Computing, BigComp, IEEE*, 2017, pp. 131–134.
- [8] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.
- [9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *International Conference on Learning Representations*, 2014.
- [10] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: *International Conference on Machine Learning, PMLR*, 2019, pp. 6105–6114.
- [11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.
- [12] S.A. Khowaja, A.G. Prabono, F. Setiawan, B.N. Yahya, S.-L. Lee, Contextual activity based healthcare internet of things, services, and people (HioTSP): An architectural framework for healthcare monitoring using wearable sensors, *Comput. Netw.* 145 (2018) 190–206.
- [13] T. Gu, L. Wang, Z. Wu, X. Tao, J. Lu, A pattern mining approach to sensor-based human activity recognition, *IEEE Trans. Knowl. Data Eng.* 23 (9) (2011) 1359–1372, <http://dx.doi.org/10.1109/TKDE.2010.184>.
- [14] H. Ma, W. Li, X. Zhang, S. Gao, S. Lu, AttnSense: Multi-level attention mechanism for multimodal human activity recognition, in: *IJCAI*, 2019, pp. 3109–3115.
- [15] H. Zhang, Z. Xiao, J. Wang, F. Li, E. Szczerbicki, A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention, *IEEE Internet Things J.* 7 (2) (2019) 1072–1080.
- [16] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [17] L. Yang, R.-Y. Zhang, L. Li, X. Xie, Simam: A simple, parameter-free attention module for convolutional neural networks, in: *International Conference on Machine Learning, PMLR*, 2021, pp. 11863–11874.
- [18] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: *Proceedings of the European Conference on Computer Vision, ECCV*, 2018, pp. 3–19.
- [19] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11534–11542.
- [21] X. Lin, L. Ma, W. Liu, S.-F. Chang, Context-gated convolution, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, Springer, 2020, pp. 701–718.
- [22] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.
- [23] M. Zeng, H. Gao, T. Yu, O.J. Mengshoel, H. Langseth, I. Lane, X. Liu, Understanding and improving recurrent networks for human activity recognition by continuous attention, in: *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, 2018, pp. 56–63.
- [24] C.A. Ronao, S.-B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, *Expert Syst. Appl.* 59 (2016) 235–244.
- [25] J. Yang, M.N. Nguyen, P.P. San, X. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: *Ijcai*, vol. 15, Buenos Aires, Argentina, 2015, pp. 3995–4001.
- [26] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
- [27] K. Kamal, Z. Yin, M. Wu, Z. Wu, Depthwise separable convolution architectures for plant disease classification, *Comput. Electron. Agric.* 165 (2019) 104948.
- [28] Z.Y. Khan, Z. Niu, CNN with depthwise separable convolutions and combined kernels for rating prediction, *Expert Syst. Appl.* 170 (2021) 114528.
- [29] C. Matsoukas, J.F. Haslum, M. Sorkhei, M. Söderberg, K. Smith, What makes transfer learning work for medical images: Feature reuse & other factors, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9225–9234.
- [30] F. Ren, W. Liu, G. Wu, Feature reuse residual networks for insect pest recognition, *IEEE Access* 7 (2019) 122758–122768.
- [31] D. Misra, T. Nalamada, A.U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3139–3148.
- [32] J. He, Q. Zhang, L. Wang, L. Pei, Weakly supervised human activity recognition from wearable sensors by recurrent attention learning, *IEEE Sens. J.* 19 (6) (2018) 2287–2297.

- [33] K. Wang, J. He, L. Zhang, Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors, *IEEE Sens. J.* 19 (17) (2019) 7598–7604.
- [34] A.K. Barbey, Network neuroscience theory of human intelligence, *Trends Cogn. Sci.* 22 (1) (2018) 8–20.
- [35] W. Gerstner, H. Sprekeler, G. Deco, Theory and simulation in neuroscience, *science* 338 (6103) (2012) 60–65.
- [36] N.K. Mehta, S.S. Prasad, S. Saurav, R. Saini, S. Singh, Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement, *Appl. Intell.* 52 (12) (2022) 13803–13823.
- [37] Y. Jiao, Z. Li, S. Huang, X. Yang, B. Liu, T. Zhang, Three-dimensional attention-based deep ranking model for video highlight detection, *IEEE Trans. Multimed.* 20 (10) (2018) 2693–2705.
- [38] B. Hariharan, J. Malik, D. Ramanan, Discriminative decorrelation for clustering and classification, in: *European Conference on Computer Vision*, Springer, 2012, pp. 459–472.
- [39] M. Aubry, B.C. Russell, J. Sivic, Painting-to-3D model alignment via discriminative visual elements, *ACM Trans. Graphics (ToG)* 33 (2) (2014) 1–14.
- [40] S.A. Hillyard, E.K. Vogel, S.J. Luck, Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence, *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 353 (1373) (1998) 1257–1270.
- [41] J.R. Kwapisz, G.M. Weiss, S.A. Moore, Activity recognition using cell phone accelerometers, *ACM SigKDD Explor. Newsl.* 12 (2) (2011) 74–82.
- [42] A. Reiss, D. Stricker, Introducing a new benchmarked dataset for activity monitoring, in: *2012 16th International Symposium on Wearable Computers*, IEEE, 2012, pp. 108–109.
- [43] D. Micucci, M. Mobilio, P. Napoletano, Unimib shar: A dataset for human activity recognition using acceleration data from smartphones, *Appl. Sci.* 7 (10) (2017) 1101.
- [44] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine, in: *Ambient Assisted Living and Home Care: 4th International Workshop, IWAAL 2012, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings 4*, Springer, 2012, pp. 216–223.
- [45] A. Ignatov, Real-time human activity recognition from accelerometer data using convolutional neural networks, *Appl. Soft Comput.* 62 (2018) 915–922.
- [46] S. Wan, L. Qi, X. Xu, C. Tong, Z. Gu, Deep learning models for real-time human activity recognition with smartphones, *Mob. Netw. Appl.* 25 (2020) 743–755.
- [47] T. Liu, S. Wang, Y. Liu, W. Quan, L. Zhang, A lightweight neural network framework using linear grouped convolution for human activity recognition on mobile devices, *J. Supercomput.* (2022) 1–21.
- [48] L. Syed, S. Jabeen, S. Manimala, A. Alsaeedi, Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques, *Future Gener. Comput. Syst.* 101 (2019) 136–151.
- [49] R. Janarthanan, S. Doss, S. Baskar, Optimized unsupervised deep learning assisted reconstructed coder in the on-nodule wearable sensor for human activity recognition, *Measurement* 164 (2020) 108050.
- [50] H. Cho, S. Yoon, Applying singular value decomposition on accelerometer data for 1D convolutional neural network based fall detection, *Electron. Lett.* 55 (6) (2019) 320–322.
- [51] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, B. Zhao, A federated learning system with enhanced feature extraction for human activity recognition, *Knowl.-Based Syst.* 229 (2021) 107338.
- [52] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (11) (2008).



Junjie Liang received the B.S. degree (2021) from Tianjin University of Technology, Tianjin, China. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include human activity recognition, computer vision, and machine learning.



Lei Zhang received the B.Sc. degree (2001) in computer science from Zhengzhou University, China, and the M.S. degree (2004) in pattern recognition and intelligent system from Chinese Academy of Sciences, China. He received the Ph.D. degree (2007) from Southeast University, China, in 2011. He was a Research Fellow (2008) with IPAM, UCLA, Los Angeles. He is currently an Associate Professor with the School of Electrical and Automation Engineering, Nanjing Normal University. His research interests include machine learning, human activity recognition and computer vision.



Can Bu received the B.S. degree from Huaiyin Institute of Technology, Huaian, China, in 2021. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include activity recognition, federated learning, and machine learning.



Guangyu Yang received the B.Sc. degree from Hefei University, Hefei, China, in 2021. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include activity recognition, computer vision, and machine learning.



Hao Wu received the B.S. degree (2001) in computer science from Zhengzhou University, He received M.S. degree (2004) and Ph.D. degree (2007) in computer science from Huazhong University of Science and Technology. Now, he is an associate professor at School of Information Science and Engineering, Yunnan University, China. He has published more than 50 papers in peer-reviewed international journals and conferences, such as JASIST, FGCS, J.Supercomputing, KBS and PUC. He has coauthored a monograph published in World Scientific. He has also served as reviewers and PC members for many venues. His research interests include service computing, information filtering and recommender systems.



Aiguo Song (Senior Member, IEEE) received the Ph.D. degree (1996) in measurement and control from Southeast University, Nanjing. He is currently a Professor with the School of Instrument Science and Engineering, Southeast University. He is the chair of the China Chapter of the IEEE Robotics and Automation Society. His current research interests include teleoperation, haptic display, the Internet telerobotics, distributed measurement systems and machine learning.