Multi-scale Context-Aware Networks Based on Fragment Association for Human Activity Recognition

Zhiqiong Wang[1,&] , Hanyu Liu[1,&] , Boyang Zhao[1] , Qi Shen[1] , Mingzhe Li[1] , Ningfeng Que[1] , Mingke Yan[1] and Junchang Xin[2,*]
[1]Medicine and Biological Information Engineering, Northeastern University
[2]Computer Science And Engineering@mail.neu.edu.cn , Northeastern University

wangzq@bmie.neu.edu.cn, {20217237, 20217299}@stu.neu.edu.cn, shenq1232019@163.com, {20217237, 20217299, 20217297, 20217253,20217285}@stu.neu.edu.cn, xinjunchang@mail.neu.edu.cn

Abstract

Sensor-based Human Activity Recognition (HAR) constitutes a key component of many artificial intelligence applications. Although deep feature extraction technology is constantly updated and iterated with excellent results, it is still a difficult task to find a balance between performance and computational efficiency. Through an in-depth exploration of the inherent characteristics of HAR data, we propose a lightweight feature perception model, which encompasses an internal feature extractor and a contextual feature perceiver. The model mainly consists of two stages. The first stage is a hierarchical multi-scale feature extraction module, which is composed of deep separable convolution and multi-head attention mechanism. This module serves to extract conventional features for Human Activity Recognition. After the feature goes through a fragment recombination operation, it is passed into the Context-Aware module of the second stage, which is based on Retentive Transformer and optimized by Dropkey method to efficiently extract the relationship between the feature fragments, so as to mine more valuable feature information. Importantly, this does not add too much complexity to the model, thereby preventing excessive resource consumption. We conducted extensive experimental validation on multiple publicly available HAR datasets.

1 Introduction

Human Activity Recognition (HAR) is an emerging research field that has attracted much attention in recent years. It aims to recognize activity information from human posture or action [Ismail *et al.*, 2023]. HAR's applications span intelligent living environments, such as motion tracking, healthcare, and human-computer interaction[Islam *et al.*, 2022].

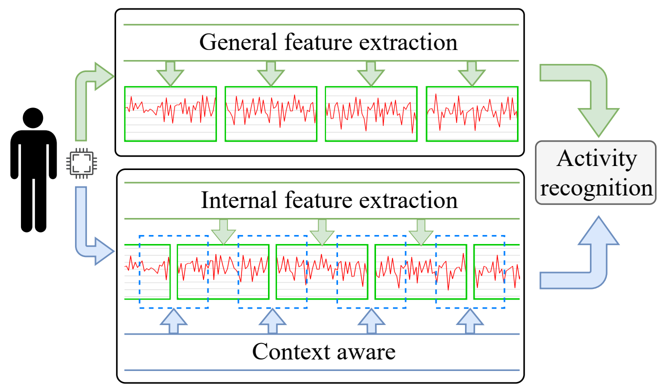

Figure 1: Comparison of our proposed method with traditional methods

Numerous studies have explored HAR, initially employing classical machine learning methods like decision trees (DT), support vector machines (SVM), random forests (RF), and naive Bayes (NB) due to their low computational complexity and suitability for smaller datasets[Wang *et al.*, 2016]. However, these methods extracted limited representative features, which constrained the classification performance. With the development of deep learning, it has become possible to automatically extract finer-grained features. Various mainstream deep neural networks such as convolutional neural networks [Zeng *et al.*, 2014] , and long short-term memory networks [Dang *et al.*, 2020a] have become important research topics in the widespread HAR scenarios, demonstrating sustained superiority. Compared to traditional machine learning methods, deep learning methods can automatically extract deep-level feature representations from sensor signals, thereby improving the accuracy of HAR[Xia *et al.*, 2020; Dua *et al.*, 2021]. Nonetheless, deep feature extraction in sensor-based HAR continues to pose several challenges:

Balancing performance and efficiency: CNNs show a strong performance in the sensor HAR advantage, small differences can effectively capture the

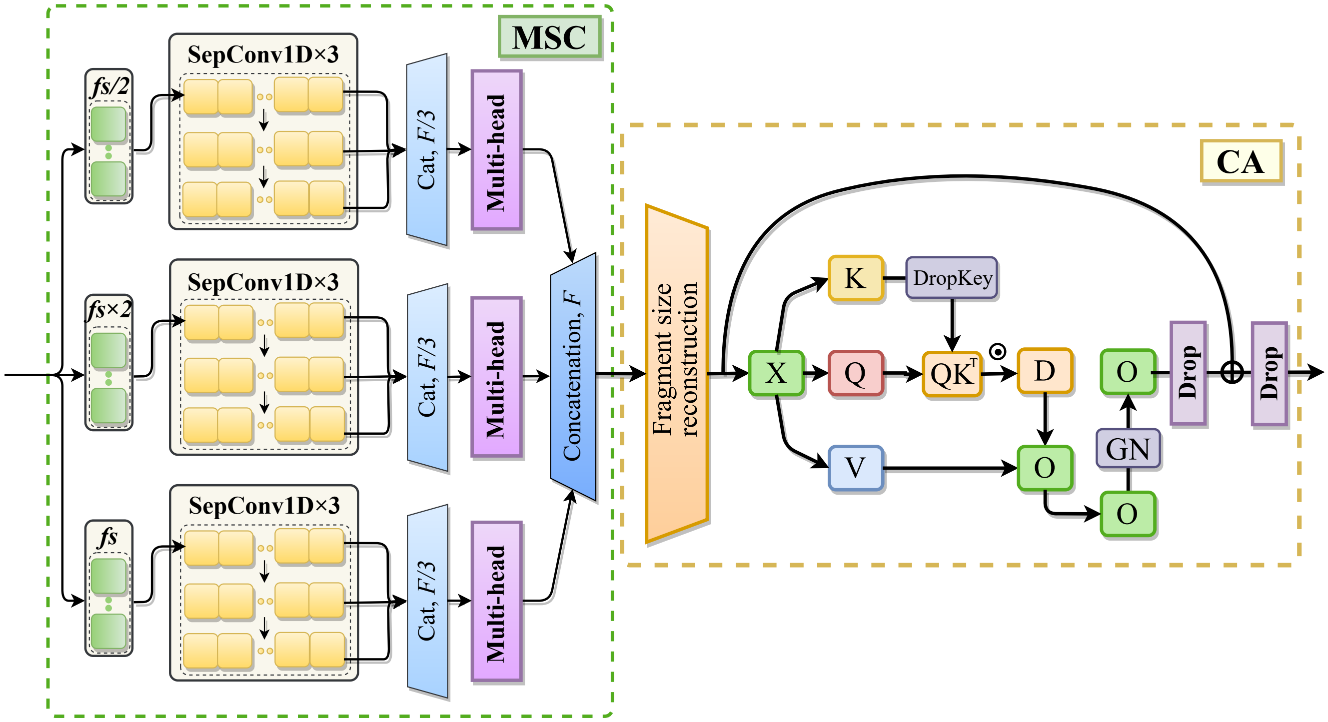

Figure 2: Flowchart of the proposed method, including the fragment association method, the internal feature extraction module MSC and the Context-Aware

activity[Gholamiangonabadi and Grolinger, 2023]. However, CNNs lack the ability to explicitly model time series data when dealing with temporal information, limiting their ability to deal with complex tasks. In contrast, time series models such as Rerrent Neural Network are better at handling time-dependent and dynamic features [Abdelrazik *et al.*, 2023]. These models are not only sensitive to temporal aspects but also have the ability to retain previous information. However, such time series models converge slowly and are prone to overfitting during training. On wearable devices, it is critical to achieve a balance between performance and efficiency in HAR systems and there is no optimal solution yet.

Distribution discrepancy: Dang et al. [Dang *et al.*, 2020b] emphasized that due to distribution differences, it is incorrect to assume that training data and test data are independent and have the same distribution in activity recognition. In sensor-based HAR, these differences mainly include distribution changes among users and changes over time [Chang *et al.*, 2020]. The difference in distribution between users is due to biological and environmental factors. For example, people walk at different speeds and stride sizes, resulting in differences between different users. Although convolutional networks are good at capturing local detail features, they are difficult to solve the above problems. In contrast, the temporal model considers the time dependence and has advantages in dealing with distribution differences in behavior recognition [Ruiz *et al.*, 2021]. However, it is still a challenge to deploy high-performance complex models on resource-constrained wearable devices.

In this paper, we have made the following contributions:

1. We propose an efficient feature extraction module, which includes an internal feature extraction module to extract intra-segment features, and a context feature awareness module to extract inter-segment relationship features.

2. We propose to use the Retentive Transformer model to capture the temporal dependence of HAR data for efficient training through a parallel representation method. And the optimization for HAR system in terms of Drop technique of self-attention.

3. Different from the previous time-dependent models, we do the correlation processing between segments of the data, so that the Context-Aware module can pay attention to the relationship between adjacent segments, so as to extract richer deep features.

The remainder of this paper is organized as follows: Section 2 reviews some relevant approaches in the field that are relevant to our work. Section 3 Outlines the current workflow and discusses the implementation details of the approach. Section 4 presents the experimental details. Finally, Section 5 shows the performance of the model in experiments, discussing the current findings.

2 Related Work

2.1 Deep Feature Extraction

In recent years, the rapid development of advanced computing resources has promoted the training of high-performance neural network models, enabling new deep learning methods to mine deeper effective features for more accurate recognition[Pramanik *et al.*, 2023; Yang *et al.*, 2020]. Qian et al. [Qian *et al.*, 2019]introduced the Distribution-Embedded Deep Neural Network, which integrates statistical features with spatial and temporal information in an end-to-end deep learning framework by incorporating an additional loss function based on Maximum Mean Discrepancy distance. Patino-Saucedo et al. [Patiño-Saucedo *et al.*, 2022] developed the Artificial and Raw Sensor fusion approach, which leverages data from multiple sensors to perform deep feature fusion encompassing counting features, averaging features, aggregated features, and raw features. Deep feature extraction often incurs substantial computational costs, and the combination of artificial and deep features emerges as a potential solution to reducing computational complexity. Ravi et al.[Ravi *et al.*, 2016] combined spectrogram features with a single CNN layer and two fully connected layers for HAR, demonstrating the feasibility of real-time applications through evaluations on four benchmark datasets.

2.2 Distribution Discrepancy

Two types of sample distribution discrepancy exist in HAR: individual user variance and temporal distribution variation. Variance among users is due to diverse movement patterns during activities, while temporal distribution changes over time with possible new activities emerging in dynamic streaming environments. The Multi-Source Unsupervised Cooperative Transfer Network (MUCT) model, proposed by Jia et al. [Jia *et al.*, 2023], addresses this distribution discrepancy through automatic feature extraction, domain adaptive alignment, and iterative use of consensus filters for improved robustness. Chen et al.[Chen *et al.*, 2019] further investigated individual differences and task consistency in human-centric sensing applications. Reducing individual differences while maintaining task consistency provides potential for accurate recognition. Furthermore, Rokne et al. [Rokni *et al.*, 2018] utilized transfer learning in personalization models, acknowledging user distribution discrepancy by fine-tuning a CNN only in the testing phase for the target user. [Liu *et al.*, 2023] proposed a fusion loss function method to optimize inter- and intra-feature problems through the idea of metric learning.

3 Methods

This paper aims to exchange less resource consumption for higher performance in sensor-based HAR systems by proposing new methods in data processing and model construction. In a standard HAR task, we first need to process raw signal data, which can be inertial sensor signals (such as three-axis acceleration and angular velocity), or ECG or EMG signals. In the data preprocessing stage, we proposed a segment association method to associate the originally scattered segment features to form a time segment feature. This can alleviate the common problem of uneven sample distribution in HAR. We denote the action segment τ^n as $x^\tau \in \mathbb{R}^T$, where T is the number of time points in a segment. Then, the feature extractor function representation comes from the input a fragment $\mathbf{f}_\tau \in \mathbb{R}^F$, where F denotes the dimension of the output feature. Then, $\mathbf{f}_\tau$ is updated to $\hat{\mathbf{f}}_\tau$ to improve the modeling context attributes. Finally, from the perspective of time series modeling, the context relationship between adjacent features is extracted by the Context-Aware module for feature extraction. Context-Aware module access to a series of characteristics $[\hat{\mathbf{f}}_\tau]_{\tau=0}^N$, and extract the relationship features between segments. In addition, we propose further optimizations for HAR data during training the model. Fig 2. shows the flow chart.

3.1 Feature Extraction within Segments

Inside the segment, we build a multi-scale hierarchical convolution module for extracting local detail features, and a multi-head attention module for extracting long-distance time-dependent features. In order to understand the relationship between time and features, we use convolution kernels of different sizes for convolution. Specifically, we use convolutional filters of size $(fs \times 2)$, (fs) and $(fs/2)$ to capture features at three scales, using one simple convolution and three separable convolutions per path. Perform three separable convolutions one after the other to get the features $x1, x2, x3$, and then concatenate them in a hierarchical way. After global pooling, the features of different scales are captured through the multi-head attention mechanism to capture the long-distance time-dependent features in the segment. Finally, the features of different scales are concated.

3.2 Context-Aware Module

Human activities exhibit significant coherence and often revolve around a single behavior for a brief duration. Consequently, it is vital to establish interrelationships between action segments. In this study, we employ a context encoder to capture these cross-segment features. Currently, similar methods have been adopted in the field of sleep monitoring. Phyo et al. [Phyo *et al.*, 2023] proposed the utilization of BiLSTM for encoding operations to capture inter-segment features and apply them to confusion classification and sleep stage representation, leading to promising outcomes. For our context module, we integrate the Retention mechanism from Retentive Network [Sun *et al.*, 2023] to handle the time correlation among fragments and enhance its structure using the DropKey method.

After obtaining the multi-scale feature representation, we introduce a Context feature awareness module called CA, incorporating the concepts from DropKey [Li *et al.*, 2023] and RetNet. The CA module employs adaptive average pooling to capture the features from the preceding multi-scale paths. Subsequently, the DK-Ret method is employed to extract temporal segment features.

3.3 Retentive Transformer with Layer-Order Key-Value Drop

We first remodel the characteristics of the context fragments through convolution processing, as MSC is introduced into the shape of (B, S, F, L) , where the characteristics of $[\mathbf{f}_\tau]_{\tau=0}^N = \{\mathbf{f}_{(-N/2)}, \mathbf{f}_{(-N/2+1)}, ..., \mathbf{f}_{(N/2-1)}, \mathbf{f}_{(N/2)}\}$,each $\mathbf{f}_\tau$ updated to $\hat{\mathbf{f}}_\tau$, access to a series of characteristics $[\hat{\mathbf{f}}_\tau]_{\tau=0}^N$, so that it can maximize the perception of the relationship between the segments, ... The retention layer is defined as follows:

$$Q = (XW_Q) \odot \Theta, \quad K = (XW_K) \odot \bar{\Theta}, \quad V = XW_V$$
$$\Theta_n = e^{in\theta}, \quad D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases} \tag{1}$$

Here, $\overline{\Theta}$ represents the complex conjugate of Θ, while $D \in \mathbb{R}^{|x|\times|x|}$ combines causal masking and exponential decay along relative distance into a single matrix . We use $h = d_{\text{model}}/d$ retention heads in each layer, where d is the head dimension.To prevent excessive feature extraction, we perform a residual join on the features before and after the DropKey-Ret operation and add dropout:

$$[\mathbf{y}_\tau]_{\tau=0}^N \leftarrow DK - Ret([\hat{\mathbf{f}}_\tau]_{\tau=0}^N) + [\hat{\mathbf{f}}_\tau]_{\tau=0}^N \tag{2}$$

The existing multi-head attention mechanism commonly incorporates the Dropout algorithm, which is often used as a regularizer in CNNs. However, using the structured drop method from CNNs is not appropriate for the multi-head self-attention model. This is because a large drop probability in the deep layer can result in the loss of high-level feature information, while a small drop probability in the shallow layer can lead to overfitting of detailed features [Li *et al.*, 2023]. To address this issue, we introduce DropKey into the time series model. For the problem of deep versus shallow features, DropKey does not perform random drops at each layer with a fixed probability. Instead, it gradually decreases the probability of drop as the number of layers deepens. The pseudo-algorithm for DropKey is presented below:

Algorithm 1: DK-Net pseudo-algorithm

Input: $\Theta_n, Q, K, V, D_{nm}$:Variables mentioned in Equation 2; $mask\_ratio$: ratio to mask;
Output: Features $\mathbf{x}$.
def Attention($Q, K, V, mask\_ratio$)
  $Attn \leftarrow$ (Scaling(Q)@K^T) $* D$
  $m_r \leftarrow$OnesLike($Attn$) $* mask\_ratio$
  $Attn \leftarrow$Softmax($Attn$+Mask(m_r) $* -1e12$)
  $x \leftarrow Attn@V$
return x

Algorithm 2: The total flow pseudo algorithm

Input: Training dataset $D = \left([\mathbf{x}_\tau]_{\tau=0}^N, [\mathbf{y}_\tau]_{\tau=0}^N\right)$; network modules: **MSC** and **CA**;
Output: Optimized parameters Θ
while network parameters not converged **do**
  Draw a sequence $([\mathbf{x}_\tau]_{\tau=0}^N, [\mathbf{y}_\tau]_{\tau=0}^N) \sim D$
  for $\tau \leftarrow 1$ *to* N **do**
    foreach i *in multi-scale* **do**
      $\mathbf{f}_{(i,\tau)} \leftarrow AAP(\mathbf{MSC}_{(i,\tau)}(x_\tau))$
    $\mathbf{f}_\tau \leftarrow$Multi-head($Cat(\sum_i \mathbf{f}_{(i,\tau)})$)
  $[\hat{\mathbf{f}}_\tau]_{\tau=0}^N \leftarrow$Update($[\mathbf{f}_\tau]_{\tau=0}^N$)
  $[\mathbf{y}_\tau]_{\tau=0}^N \leftarrow$DK-Ret$_{(i,\tau)}$($[\hat{\mathbf{f}}_\tau]_{\tau=0}^N$) $+ [\hat{\mathbf{f}}_\tau]_{\tau=0}^N$
  $\mathscr{L} \leftarrow \frac{1}{N}\sum$WCE$(\mathbf{y}_\tau, \hat{\mathbf{y}}_\tau) + \lambda\cdot$WCS$(\mathbf{y}_\tau, \hat{\mathbf{y}}_\tau)$
  $\Theta \leftarrow$Adam($\mathscr{L}$)

3.4 Optimization

To optimize the tunable parameters of MSC and DK-Ret, we develop a classification learning task. Considering that HAR data basically have class imbalance problem, we also use WCE function and class weighted cosine similarity (WCS) objective function [Phyo *et al.*, 2023]:

$$\text{WCS}\left(\mathbf{y}_\tau, \hat{\mathbf{y}}_\tau\right) = -w_c \sum_{\tau=0}^{N} \left(1 - \cos\left(\mathbf{y}_\tau, \hat{\mathbf{y}}_\tau\right)\right) \tag{3}$$

where $\cos(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot \mathbf{w}/\|\mathbf{v}\| \cdot \|\mathbf{w}\|$ is the cosine similarity operation and λ denotes a scaling hyperparameter. Pseudo algorithms for training all networks in the framework are given in Algorithm 2.

4 Experiment

The experiments are carried out on the Kaggle platform, and we choose NVIDIA P100 and the default configuration. The experimental design consists of two main parts: ablation experiments and a comparison of related work.

4.1 Datasets

This section offers a comprehensive elucidation of the conducted experiments and specificities associated with them. The pivotal component of these investigations were four distinct datasets, each characterized by a unique method of data acquisition. Data were accumulated either through multiple sensor nodes or via smartphones, carried by participants as they engaged in various activities across different contexts. The datasets encompassed within our research include OPPORTUNITY, PAMAP2, and UCI-HAR, effectively forming a composite of multimodal HAR data. In addition, we incorporated the WISDM dataset that was compiled using a tri-axial accelerometer. In order to ensure an objective evaluation of our methodology, several pertinent aspects of these employed datasets have been outlined below.

Datasets	PAMAP2	WISDM	OPPO.	UCI-HAR
Sensor	40	3	72	9
Subject	9	29	12	30
Class	12	6	18	6
Window Size	171	90	113	128
Sequence	4	8	32	8
Batch Size	512	512	256	512
Lr	0.001	0.001	0.001	0.001
Epoch	50	50	40	50

Table 1: Dataset Details

PAMAP2 [Reiss and Stricker, 2012]: The PAMAP2 Physical Activity Monitoring dataset is public available at UCI repository, which contains 18 different physical activities The dataset was collected from 9 subjects who wore 3 wireless Inertial Measurement Units.

WISDM [Kwapisz *et al.*, 2011]: A public dataset provided by the Wireless Sensor Data Mining Laboratory, containing 6 data attributes: user, activity, timestamp, x, y, z. Twenty-nine volunteers were recruited to perform a specific set of activities.

OPPORTUNITY[Chavarriaga *et al.*, 2013]: Realistic daily life activities of 12 subjects in a sensor-enriched environment were recorded. In total, 15 networked sensor systems, including 72 sensors in 10 modalities, are integrated on the environment and the body.

UCI-HAR [Ignatov, 2018]: It contains sensor recordings from 30 subjects who were asked to wear a waist-mounted smartphone to perform six activities of daily living (ADLs). The triaxial acceleration and triaxial angular velocity signals were collected at a sampling rate of 50Hz during data acquisition.

4.2 Evaluation Metrics

To evaluate the performance of the proposed model for HAR, the following metrics were used for evaluation generally.

$$\begin{aligned}
\text{Accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \\
\text{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\
\text{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
\text{F1-macro} &= \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \\
\text{F1-weighted} &= \sum_i \frac{2\times\omega_i\times(\text{Precision}_i\times\text{Recall}_i)}{\text{Precision}_i+\text{Recall}_i}
\end{aligned} \tag{4}$$

where TP and TN are the number of true and false positives, respectively, and FN and FP are the number of false negatives and false positives. ω_i is the proportion of samples of class i.

Network	Accuracy	F1-macro	F1-weight	Time
MSC	96.85%	95.44%	96.89%	3m49s
	90.12%	90.56%	91.43%	1m57s
MSC-CA	97.93%	96.78%	97.95%	5m 1s
	96.89%	96.43%	97.02%	2m28s

Table 2: Ablation Experiment

5 Results & Discussion

5.1 Ablation Experiment

We conduct experiments to evaluate the performance and effectiveness of different combinations of modules. Second, the performance and efficiency of the MSC-CA model when using only MSC modules or both MSC and CA modules are verified in stages to ensure its plausibility. Two datasets, WISDM and PAMAP2, were selected for ablation experiments, and the specific information is in Table 2.

The MSC model solely captures internal segment information, but demonstrates effective performance and rapid convergence. In comparison, the MSC-CA model, which incorporates relationship features between segments, exhibits superior performance, surpassing the MSC model across all three evaluation indicators. On the WISDM dataset, we achieve an average performance lead of 1%. This advantage becomes more pronounced on PAMAP2, where we observe a lead of 6%. This disparity can be attributed to the increased complexity and a multitude of actions in PAMAP2, making it more challenging to identify confusing categories. Moreover, in terms of performance and efficiency, the addition of the Context-Aware module in MSC-CA incurs minimal overhead in time when compared to MSC. By simultaneously extracting internal segment features and inter-segment relationship features, our approach outperforms the method of solely extracting internal segment features, thus achieving a better tradeoff between effectiveness and efficiency. Furthermore, regarding the distribution differences, depicted in Fig 3 and 4, we present the confusion matrices of MSC-CA for WISDM and PAMAP2. The left figure reveals some confusion between the ”Downstairs” and ”Upstairs” classes, while the right figure indicates confusion primarily caused by the scarcity of samples. Categories with smaller motion amplitudes, such as ”Sitting” and ”Standing,” do not appear, which stems from distribution disparities. Notably, MSC-CA exhibits approximately 5% higher accuracy for ”downstairs” and ”upstairs” compared to MSC, highlighting the significant enhancement brought by the Context-Aware module.

5.2 Comparison with Existing Work

We further investigate the performance and efficiency of our model. We use all four datasets mentioned in the article for detailed testing, and the test metrics include accuracy and time. Fig 6 and Table 3 visually compares the performance

Model	PAMAP2		WISDM		OPPORTUNITY		UCI-HAR	
CNN [Zeng *et al.*, 2014]	90.86%	0m24s	93.31%	0m37s	82.15%	0m39s	92.39%	0m14s
LSTM[Dang *et al.*, 2020a]	89.71%	0m43s	96.71%	1m41s	81.65%	0m68s	95.52%	0m12s
LSTM-CNN [Xia *et al.*, 2020]	90.48%	0m48s	95.90%	2m28s	77.64%	1m36s	97.01%	0m32s
CNN-GRU [Dua *et al.*, 2021]	90.10%	0m25s	94.95%	1m18s	79.85%	0m44s	95.11%	0m17s
SE-Res2Net [Gao *et al.*, 2019]	90.91%	1m32s	95.52%	8m47s	82.15%	2m58s	96.60%	1m23s
ResNeXt[Mekruksavanich *et al.*, 2022]	90.52%	18m24s	96.67%	22m44s	79.15%	9m21s	96.38%	3m48s
Gated-Res2Net [Yang *et al.*, 2020]	91.81%	2m46s	97.02%	8m36s	81.51%	3m22s	96.31%	2m51s
Rev-Attention [Pramanik *et al.*, 2023]	89.90%	2m26s	97.46%	6m14s	83.77%	2m18s	95.53%	2m28s
MSC-CA	**96.89%**	**2m28s**	**97.93%**	**4m 52s**	**84.63%**	**1m23s**	**96.52%**	**2m44s**

Table 3: Comparison with Existing Work

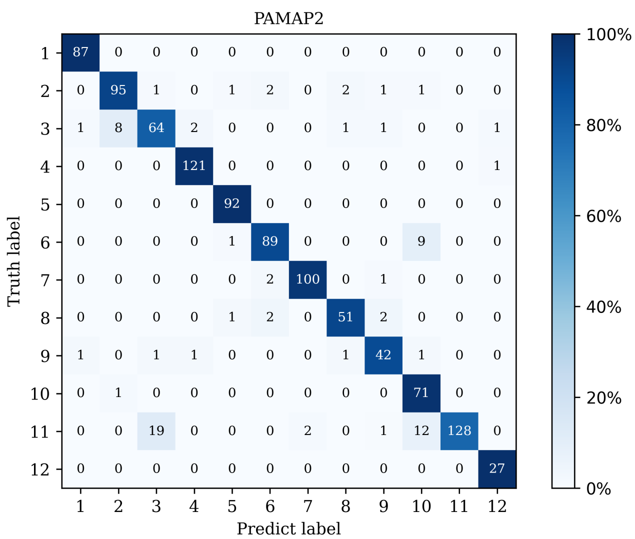

Figure 3: Confusion matrix of the model on PAMAP2

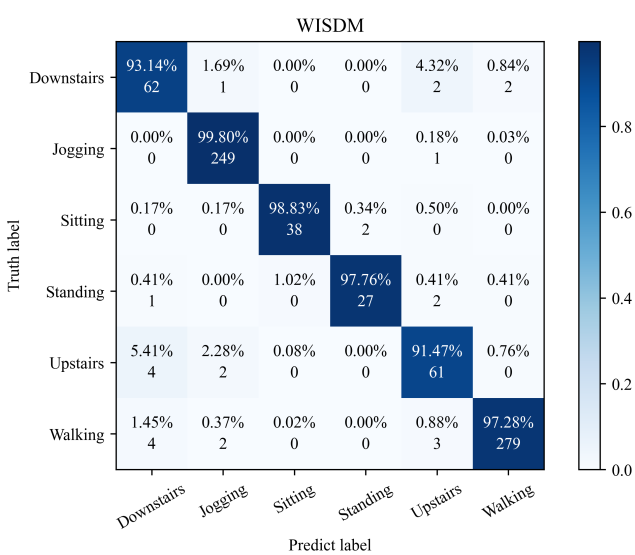

Figure 4: Confusion matrix of the model on WISDM

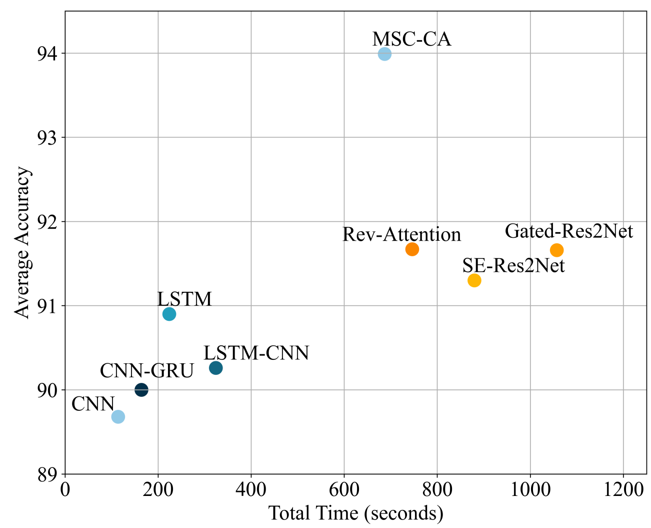

Figure 5: Scatter plot of model performance versus efficiency

and efficiency between MSAP-DM and advanced models.

In terms of performance, both the UCI-HAR and WISDM datasets demonstrate high accuracy across all models. MSC-CA exhibits a notable advantage over the most advanced HAR model, achieving accuracies of 96.52% and 97.93%, respectively, representing a significant 1% improvement. This indicates the effectiveness of the Context-Aware module even for simpler datasets. However, on more complex classification datasets like PAMAP2 and OPPORTUNITY, MSC-CA outperforms the others with even greater margins, attaining accuracies of 96.89% and 84.63%, respectively. These results are approximately 5% and 2% better than the next best models for their respective datasets. Notably, all models achieve remarkably high accuracy ($\geq$98%) in general categories. Therefore, MSC-CA demonstrates superior classification performance, particularly for challenging and confusing categories.

In terms of efficiency, Fig 5 shows the relationship between the average accuracy and the total time of the model on the four datasets, and it can be seen that even without the time series network module, the most advanced model

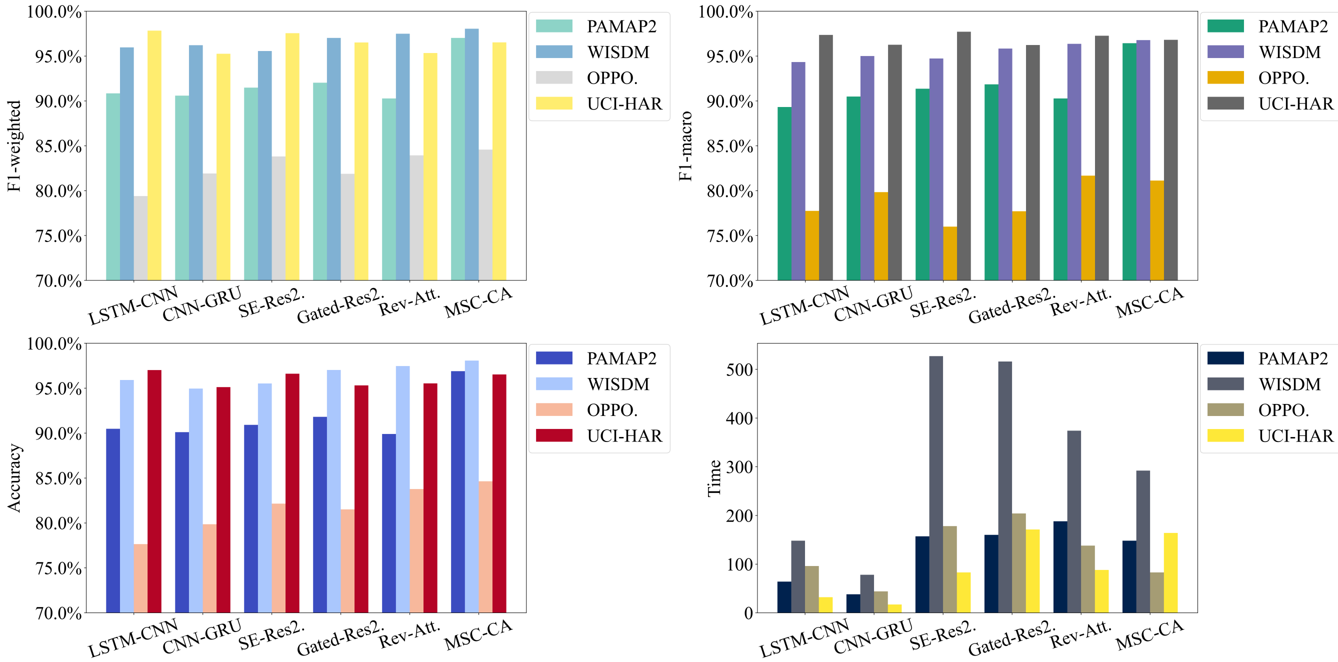

Figure 6: Model performance and efficiency analysis

consumes more time than the classical model, but the performance is significantly better than the classical models such as CNN, LSTM, CNN-GRU. Although there is a big gap in time consumption, in most cases, the performance improvement brought by advanced models is worth it. After adding the time series module (CA), our MSC-CA model maintains comparable time consumption with the state-of-art performance models, and especially obtains very excellent performance on PAMAP2 and OPPORTUNITY, whose increased time consumption brings much higher benefits than other state-of-art models.

In general, we believe that the MSC-CA model has reached a new height in the trade-off between performance and efficiency. We have a good control of model size and performance. Compared with other model, the performance improvement value we get is much higher than the cost. In addition, our model shows strong performance for complex datasets, which helps the field of wearable HAR to continue to develop towards high-performance complex tasks in the future.

6 Conclusion

In this paper, we deeply study the temporal correlation of HAR data and propose a lightweight network model with Powerful feature extraction ability. MSC-CA consists of an internal feature extraction module and a Context-Aware module. The model effectively captures the correlation information between segments while maintaining lightweight, so as to alleviate the problems of efficiency and distribution difference in HAR. Our additional processing of the data also allows the model to better capture the connections between segments. We experimentally verify the effectiveness and frontier of the proposed method, which provides a new idea for the application of sensor-based HAR. We will explore more realistic methods in the future and plan to conduct real-machine deployment on embedded devices to prove the practical usability of the method.

Acknowledgements

This work was supported by National Natural Science Foundation of China (62072089); Fundamental Research Funds for the Central Universities of China (N2116016, N2104001 and N2019007); National Training Program of Innovation and Entrepreneurship for Undergraduates(202310145023).

Contribution Statement

*This is our corresponding author. &These authors contributed equally to this work and should be considered co-first authors. Hanyu Liu & Boyang Zhao: Conceptualisation (equal), Methodology (lead), Writing - Original Draft (lead), Writing - Review & Editing (lead). Qi Shen: Conceptualisation (lead), Formal Analysis (lead), Data Curation (equal), Writing - Review & Editing (lead). Mingzhe Li & Ningfeng Que: Investigation (lead), Formal Analysis (lead), Data Curation (equal), Writing - Review & Editing (equal). Mingke Yan: Methodology (supporting), Validation (equal), Data Curation (equal). Zhiqiong Wang & Junchang Xin: Conceptualisation (lead), Resources (lead), Supervision (lead).

References

[Abdelrazik *et al.*, 2023] Mostafa A Abdelrazik, Abdelhaliem Zekry, and Wael A Mohamed. Efficient hybrid algorithm for human action recognition. *Journal of Image and Graphics*, 11(1):72–81, 2023.

[Chang *et al.*, 2020] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–30, 2020.

[Chavarriaga *et al.*, 2013] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[Chen *et al.*, 2019] Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. Distributionally robust semi-supervised learning for people-centric sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3321–3328, 2019.

[Dang *et al.*, 2020a] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

[Dang *et al.*, 2020b] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

[Dua *et al.*, 2021] Nidhi Dua, Shiva Nand Singh, and Vijay Bhaskar Semwal. Multi-input cnn-gru based human activity recognition using wearable sensors. *Computing*, 103:1461–1478, 2021.

[Gao *et al.*, 2019] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[Gholamiangonabadi and Grolinger, 2023] Davoud Gholamiangonabadi and Katarina Grolinger. Personalized models for human activity recognition with wearable sensors: deep neural networks and signal processing. *Applied Intelligence*, 53(5):6041–6061, 2023.

[Ignatov, 2018] Andrey Ignatov. Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, 62:915–922, 2018.

[Islam *et al.*, 2022] Md. Milon Islam, Sheikh Nooruddin, Fakhri Karray, and Ghulam Muhammad. Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects. *Computers in Biology and Medicine*, 149:106060, 2022.

[Ismail *et al.*, 2023] Walaa N. Ismail, Hessah A. Alsalamah, Mohammad Mehedi Hassan, and Ebtesam Mohamed. Auto-har: An adaptive human activity recognition framework using an automated cnn architecture design. *Heliyon*, 9(2):e13636, 2023.

[Jia *et al.*, 2023] Qi Jia, Jing Guo, Po Yang, et al. A holistic multi-source transfer learning approach using wearable sensors for personalized daily activity recognition. *Complex & Intelligent Systems*, 2023, 2023.

[Kwapisz *et al.*, 2011] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[Li *et al.*, 2023] Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22700–22709, 2023.

[Liu *et al.*, 2023] Hanyu Liu, Boyang Zhao, Chubo Dai, Boxin Sun, Ang Li, and Zhiqiong Wang. Mag-res2net: A novel deep learning network for human activity recognition. *Physiological Measurement*, 44(11):115007, 2023.

[Mekruksavanich *et al.*, 2022] Sakorn Mekruksavanich, Ponnipa Jantawong, and Anuchit Jitpattanakul. A deep learning-based model for human activity recognition using biosensors embedded into a smart knee bandage. *Procedia Computer Science*, 214:621–627, 2022.

[Patiño-Saucedo *et al.*, 2022] Janns Alvaro Patiño-Saucedo, Paola Patricia Ariza-Colpas, Shariq Butt-Aziz, Marlon Alberto Piñeres-Melo, José Luis López-Ruiz, Roberto Cesar Morales-Ortega, and Emiro De-la-hoz Franco. Predictive model for human activity recognition based on machine learning and feature selection techniques. *International Journal of Environmental Research and Public Health*, 19(19), 2022.

[Phyo *et al.*, 2023] Jaeun Phyo, Wonjun Ko, Eunjin Jeon, and Heung-Il Suk. Transsleep: Transitioning-aware attention-based deep neural network for sleep staging. *IEEE Transactions on Cybernetics*, 53(7):4500–4510, 2023.

[Pramanik *et al.*, 2023] Rishav Pramanik, Ritodeep Sikdar, and Ram Sarkar. Transformer-based deep reverse attention network for multi-sensory human activity recognition. *Engineering Applications of Artificial Intelligence*, 122:106150, 2023.

[Qian *et al.*, 2019] Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. A novel distribution-embedded neural network for sensor-based activity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5614–5620. International Joint

Conferences on Artificial Intelligence Organization, 7 2019.

[Ravi *et al.*, 2016] Daniele Ravi, Charence Wong, Benny Lo, and Guang-Zhong Yang. Deep learning for human activity recognition: A resource efficient implementation on low-power devices. In *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*, pages 71–76. IEEE, 2016.

[Reiss and Stricker, 2012] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[Rokni *et al.*, 2018] Seyed Ali Rokni, Marjan Nourollahi, and Hassan Ghasemzadeh. Personalized human activity recognition using convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[Ruiz *et al.*, 2021] Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021.

[Sun *et al.*, 2023] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

[Wang *et al.*, 2016] Zhelong Wang, Donghui Wu, Jianming Chen, Ahmed Ghoneim, and Mohammad Anwar Hossain. A triaxial accelerometer-based human activity recognition via eemd-based features and game-theory-based feature selection. *IEEE Sensors Journal*, 16(9):3198–3207, 2016.

[Xia *et al.*, 2020] Kun Xia, Jianguang Huang, and Hanyu Wang. Lstm-cnn architecture for human activity recognition. *IEEE Access*, 8:56855–56866, 2020.

[Yang *et al.*, 2020] Chao Yang, Mingxing Jiang, Zhongwen Guo, and Yuan Liu. Gated res2net for multivariate time series analysis. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[Zeng *et al.*, 2014] Ming Zeng, Le T. Nguyen, Bo Yu, Ole J. Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th International Conference on Mobile Computing, Applications and Services*, pages 197–205, 2014.