Northumbria Research Link

Citation: Hamad, Rebeen, Yang, Longzhi, Woo, Wai Lok and Wei, Bo (2023) ConvNet-based performers attention and supervised contrastive learning for activity recognition. Applied Intelligence, 53 (8). pp. 8809-8825. ISSN 0924-669X

Published by: Springer

URL: https://doi.org/10.1007/s10489-022-03937-y <https://doi.org/10.1007/s10489-022-03937-y>

This version was downloaded from Northumbria Research Link: https://nrl.northumbria.ac.uk/id/eprint/49498/

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: http://nrl.northumbria.ac.uk/policies.html

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

Northumbria University NEWCASTLE

University Library

Applied Intelligence (2023) 53:8809–8825
https://doi.org/10.1007/s10489-022-03937-y

ConvNet-based performers attention and supervised contrastive learning for activity recognition

Rebeen Ali Hamad[1] · **Longzhi Yang**[1] · **Wai Lok Woo**[1] · **Bo Wei**[2]

Accepted: 24 June 2022 / Published online: 3 August 2022
© The Author(s) 2022

Abstract

Human activity recognition based on generated sensor data plays a major role in a large number of applications such as healthcare monitoring and surveillance system. Yet, accurately recognizing human activities is still challenging and active research due to people's tendency to perform daily activities in a different and multitasking way. Existing approaches based on the recurrent setting for human activity recognition have some issues, such as the inability to process data parallelly, the requirement for more memory and high computational cost albeit they achieved reasonable results. Convolutional Neural Network processes data parallelly, but, it breaks the ordering of input data, which is significant to build an effective model for human activity recognition. To overcome these challenges, this study proposes causal convolution based on performers-attention and supervised contrastive learning to entirely forego recurrent architectures, efficiently maintain the ordering of human daily activities and focus more on important timesteps of the sensors' data. Supervised contrastive learning is integrated to learn a discriminative representation of human activities and enhance predictive performance. The proposed network is extensively evaluated for human activities using multiple datasets including wearable sensor data and smart home environments data. The experiments on three wearable sensor datasets and five smart home public datasets of human activities reveal that our proposed network achieves better results and reduces the training time compared with the existing state-of-the-art methods and basic temporal models.

Keywords Activity recognition · Contrastive learning · Class imbalanced problem · Temporal evaluation · Attention mechanism · Sensor data

1 Introduction

Sensors from smart home environments and wearable objects generate a large amount of valuable data used for different applications including human activity recognition (HAR). Smart home environments based on equipped sensors are designed for ambient assisted living to unobtrusively track human activities [12]. Further, wearable sensors have been also used to gather customized data about users' habits. Wearable sensors can be embedded into different objects such as mobile, clothes, belts, wristwatches, or glasses which can be worn to record users' movement with the aim of HAR [19]. Moreover, wearable and smart home sensors can record perceived information to sufficiently detect the ambulatory and postural activities [5, 29].

HAR is an active and challenging research field in ubiquitous computing to understand human activities, which plays a significant role in several applications in the fields of healthcare monitoring [30], security surveillance systems [27], and resident situation assessment [21]. HAR, as one of the important applications of healthcare monitoring from sensors data, is used to monitor and track vulnerable people [25]. However, human activities are highly diverse due to different sensor readings and even the same subject tends to perform an activity in different ways. Also, the

✉ Rebeen Ali Hamad
rebeen.hamad@northumbria.ac.uk

Longzhi Yang
longzhi.yang@northumbria.ac.uk

Wai Lok Woo
wailok.woo@northumbria.ac.uk

Bo Wei
bo.wei@lancaster.ac.uk

[1] Department of Computer and Information Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, UK

[2] School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, UK

intrinsic characteristic of categories denoting daily human activities is inherently imbalanced, and hence building a robust machine learning model for HAR is challenging. Moreover, occasionally generated data by sensors could be noisy which adds extra challenges and ambiguity to the interpretation of human activities [13].

Deep learning models are widely employed in different applications of computer vision, audio recognition, and natural language processing. Furthermore, deep learning approaches have improved HAR systems based on sensors generated data and show promising results. Since mostly HAR problems are formed as a sequential learning [22], Recurrent Neural Network (RNN) as a type of sequential learning and its variations particularly Long Short-Term Memory (LSTM) have demonstrated satisfying and state-of-the-art performance [25]. LSTM integrated models are commonly used and increase the performance of HAR systems, however, LSTM requires a large amount of memory and high computational capacity for its memory cells and gating mechanism in learning to process temporal sequential contextual information [13]. Further, LSTM models process timesteps of sensors temporal data sequentially because processing any timestep requires the outcomes of the previous timesteps [2, 42]. Convolutional Neural Network (ConvNet) is employed to extract the temporal contextual information for HAR systems from sensors data [13, 36]. Even though the training of one dimensional (1D) ConvNet models is remarkably faster than LSTM due to the nonexistence of recurrent settings, LSTM models show better performance than 1D ConvNet for HAR systems. Furthermore, 1D ConvNet models are not sensitive to the order of the sensors sequential data which is crucial for HAR due to processing sensors sequential temporal data in parallel.

The self-attention technique is used to focus more on important timesteps of the feature maps by computing similarity scores for all timesteps [42]. However, computational and memory requirements of the self-attention technique are quadratic with the length of the input sensor sequential data which leads to slow learning and occupying more memory.

To overcome the above challenges, we propose the causal ConvNet based on performers-attention and supervised contrastive learning. The proposed network improves the results of the HAR systems in sensors generated data. In addition, the proposed method also accelerates the learning process compared to the existing methods. Causal convolution [2, 31] is adopted to avoid violating the ordering timesteps of the input datasets, which is crucial in HAR systems. Performers-attention [6] which scales linearly with the input sequence length is proposed to reduce the computation and memory cost compared to the self-attention mechanism for HAR systems. Moreover, supervised contrastive learning is adopted to learn a good representation from the input sensors data that supports classifiers to gain useful information [3, 20]. Due to integrating supervised contrastive learning, the proposed network has two learning stages. The network learns a good representation of human activities in the first stage to learn a more accurate classifier in the second stage. Further, in the first stage, the supervised contrastive loss function is applied to learn the representation of human activities which is further propagated through a projection network. In the second stage, a linear classifier is trained on top of the frozen representations while the projection network is discarded. The two stages of learning prepare a discriminative representation that renders a more accurate classifier [20].

Moreover, due to the diversity of human activity recognition which leads to generating long-tailed datasets with skewed class distributions, often classifiers tend to be more biased towards majority classes and misclassify minority classes. To address this limitation, the focal loss function [23] based upon the effective number of samples [7] is proposed by assigning higher weights to hard-classified examples to sufficiently learn minority classes. The focal loss function is conducted in the second stage to learn a linear classifier for HAR. The proposed network is evaluated on eight benchmark HAR datasets and compared with the existing state-of-the-art methods. The experimental results demonstrate that our proposed network can obtain better results compared with the existing state-of-the-art methods. An ablation study is carried out to demonstrate the contribution of each of the components (performer attention, supervised contrastive learning: two stages learning, causal convolution, and focal loss) of the proposed network.

To summarise, we propose a causal ConvNet-based performers-attention and supervised contrastive learning to increase the accuracy of HAR systems and accelerate the learning process. The main components of the proposed network are described below:

i. The performers-attention is adopted to effectively expose significant timesteps that involve human activities.
ii. Supervised contrastive learning within the network is proposed to render expressive representations that help the classifier to accurately and easily recognize human activities.
iii. Causal convolutions as part of the network are proposed to maintain the ordering of sensor data which is important for systems of HAR by preventing information flow from future to past.
iv. The focal loss function based on the effective number of samples is proposed to down-weights well-classified examples and focus on hard-classified examples.

The remainder of this paper is structured as follows. The related works is reviewed in Section 2. A background for this study is provided in Section 3. The details of the proposed network is described in Section 4. Section 5 reports evaluations of the experimental setup. Finally, Section 6 concludes the paper.

2 Related works

Deep learning models have shown a significant breakthrough with appreciable performance on different HAR benchmark datasets [11]. Moreover, deep learning models are used not only in the form of single model learning but also joint models learning to address class imbalanced problems and improve HAR systems [15]. Since HAR is a sequential classification problem, recurrent network-based architectures, i.e., RNN and LSTM, have shown satisfying results. HAR based on RNN is conducted to recognize human activities from sensors data [8]. Although the results achieved based on RNN are reasonable, RNN cannot prevent gradient vanishing and exploding problems in processing long input sequences [25]. LSTM [17] was developed to prevent the occurrence of exploding problems and vanishing gradients using multiple switch gates. LSTM can process long-term dependencies of temporal sequential data including HAR systems. Several studies have used LSTM to model human activities from sensor data [11, 25, 33]. Moreover, LSTM is not only used alone or in ensemble form [25] to model human activities but also combined with ConvNet vertically or parallelly to process long-term dependencies and enhance HAR systems [11]. ConvNet is used for HAR systems to dispense recurrent architectures, make the learning phase faster, and process sequential temporal human activities in parallel [36].

Self-attention mechanism [42] is used with recurrent-based networks and ConvNet to focus more on the most relevant time steps and increase the accuracy of HAR systems. Hybrid ConvNet and LSTM with self-attention mechanism are used for HAR using reinforcement learning from wearable sensors [16]. Due to this hybrid method trained based on reinforcement learning, large computing resources for the training phase are required. Furthermore, the self-attention mechanism is appended to the Convolutional LSTM model for HAR to pay more attention to informative timesteps from temporal sequential wearable sensor data [37]. The self-attention mechanism is employed in further studies of HAR based on wearable sensors data [4, 24]. Recurrent network architecture from these methods leads to a delay in the training process. Moreover, these methods are built only based on wearable sensors data for HAR systems. Moreover, the DeepConvLSTM method is suggested for HAR based on sensor data from smart homes [26]. Due to the recurrent setting in this method, parallelization in processing the input sequence is restricted which makes this model computationally expensive and occupies more memory. Further, this model is only compared to bidirectional LSTM and evaluated on three smart home datasets. ConvNet based on dual attention is proposed that entirely dispense recurrent settings for activity recognition, however, the proposed model is only evaluated on wearable sensor datasets [9].

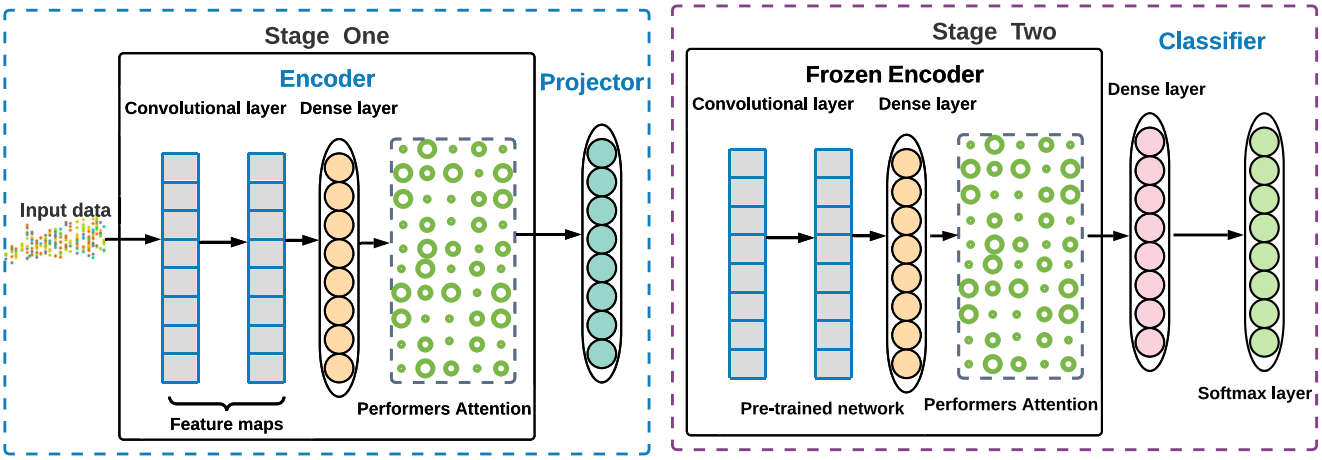
Despite the effectiveness of self-attention for HAR, computation and memory cost of the self-attention technique scales quadratically with the length of the data which delays the learning process. To remedy these limitations and enhance the performance of HAR systems from sensors data, we propose causal ConvNet-based performers-attention and supervised contrastive learning. This is because firstly the performers-attention mechanism linearly scales with the length of the sensor input data which makes the learning process faster. Secondly, supervised contrastive learning increases the performance of the proposed network by replacing one stage learning with two stages of learning where the first stage is representation learning and the second stage is classifier learning.

3 Background

3.1 Self-attention

Self-attention is a powerful mechanism that computes correlation scores for all pairs of the samples in the input sensor data. The self-attention mechanism is introduced and exploited by the Transformer architecture to process sequential data in parallel [42]. To make the model pay extra attention to the essential time steps in modelling HAR from the temporal sensor representation, the self-attention technique is employed in the training phase. Attention technique has the following learned components: query Q, key K, and values V. The dimension size of query Q and key K is d_k, where the dimension size of V is d_v [42]. The complexity of the self-attention mechanism with the length of the input temporal sequence scales quadratically which increases model learning time and requires more memory. This is the limitation of the self-attention mechanism which is addressed in Section 4.2. The attention matrix implementation is shown in (1).

$$\boldsymbol{Z}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{1}{\sqrt{d_k}}\boldsymbol{Q} \cdot \boldsymbol{K}^T\right)\boldsymbol{V}. \tag{1}$$



(a) Representation learning (b) Classifier learning

Fig. 1 Proposed network

3.2 Contrastive learning

Contrastive learning [10] has been employed for supervised and unsupervised learning as an objective function [20, 28]. The purpose of contrastive learning is to learn $f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^d$ (a parametric function) that able to map an input data x to a feature map ($f_\theta(x) \in \mathbb{R}^d$ with $d < D$) so that a cosine distance as a distance measure can project a high-dimensional input space with complex similarities to a low-dimensional feature latent embedded space. Generally, contrastive learning aims to learn representations by mapping input data to a feature space where similar examples are close together and dissimilar examples are far apart [10]. Hence, contrastive learning increases both compactnesses of intra-classes and separability of inter-classes which lead to rendering a better classifier. Moreover, learning representations of the input data support classifiers to easily extract useful information to properly distinguish categories [3]. The supervised contrastive learning [20] maps the encoded normalized samples belonging to the same class close together in embedding space and simultaneously pushing apart clusters of samples from different categories.

4 Proposed network

The proposed network is built using causal 1D ConvNet with the performers-attention based on supervised contrastive learning. The proposed method takes the minority classes from the input datasets into consideration using the focal loss function with an effective weighting samples technique as described in Section 4.1. The causal convolutions component in the proposed network is used to avoid information flow from future to past by processing results at time t based on solely the convolutions of the time steps of the temporal data from time t and earlier in the previous layer. Therefore, predicting time steps at time t cannot rely on any of the future time steps from the sensor sequential data. This helps the proposed network to maintain the ordering

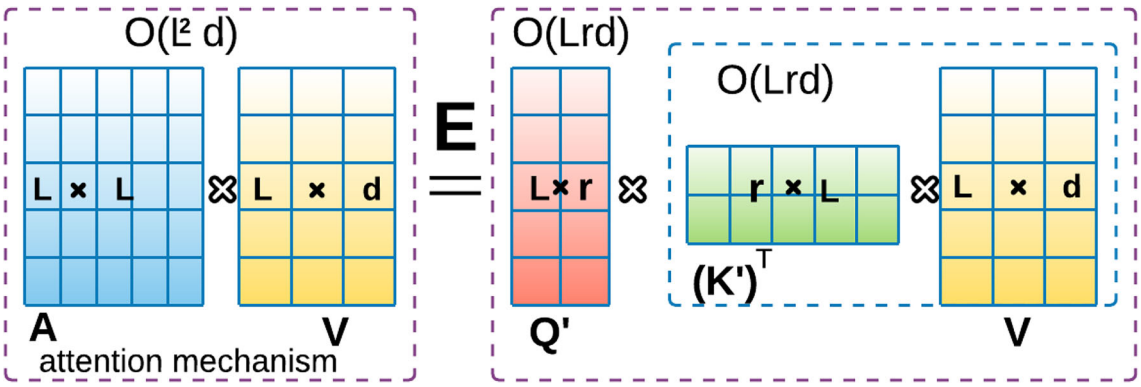


Fig. 2 Approximation of the regular attention mechanism AV via random feature maps. Dashed blocks show the order of computation with corresponding time complexities [6]

of the temporal data [31] which is significant for HAR systems [13]. Moreover, the details of the performers-attention are provided in Section 4.2. Figures 1 and 2 presents the structure of the proposed network and the two stages of learning in which the representation learning uses supervised contrastive loss function and the classifier learning uses the focal loss function. More details about supervised contrastive learning and both learning stages are provided in Section 4.3.

4.1 Focal loss

The focal loss [23] is introduced to address the imbalanced class problem between background and foreground classes during training in one stage object detection scenario. The focal loss is designed to down-weight well-classified examples and focuses on hard-classified examples. The loss value of hard-classified examples is much higher compared to the loss values of the well-classified examples by a classifier using the focal loss function. Since the focal loss focuses more on a sparse set of hard-classified samples, hence the focal loss is used in our proposed network to improve the learning of minority classes in HAR systems. The focal loss function is shown in (2).

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{2}$$

4.2 Generalized kernelizable attention

The complexity of the self attention mechanism with the length of the input temporal sequence scales quadratically which increases model learning time and requires more memory. This is the limitation of the self-attention mechanism. To address this limitation, we adopt performers-attention [6] as an efficient attention mechanism whose complexity scales linearly with the size of an input sequence L. The performers uses a Fast Attention Via Positive Orthogonal Random Features (FAVOR+) algorithm and substitutes Transformer self-attention by generalized kernelizable attention. The FAVOR+ algorithm is used to estimate the regular softmax attention by random feature map decompositions. Hence the core idea of the performers is to decompose the attention matrix into a matrix product. This algorithm leverages positive orthogonal random features to approximate softmax attention kernels with provable accuracy and $O(N)$ for both computational and space complexity [6]. Previous attention mechanisms such as sparsity and low-rankness relied on structural assumptions for the attention matrix without approximating the original softmax function. Generalized kernelizable attention can make the model process longer input sequences and train faster compared to previous attention mechanisms. The aim of using generalized kernelizable attention and FAVOR+ is to approximate the softmax and choose the order of computation of the matrices of (1).

4.3 Supervised contrastive learning

In this study, supervised contrastive learning (SCL) is used to build a model for HAR that outperforms the state-of-the-art HAR methods. The proposed method based on SCL consists of two stages of learning. In the first stage, two components are trained which are encoder and projection networks. The first stage learns representations used in the second learning stage to build a robust and accurate classifier for HAR systems. The details of the first stage are as follows:

1. Encoder network $E(\cdot)$ maps temporal input sequential data x to a representation vector $r = E(x) \in R^{D_E}$ where $D_E = 512$. The encoder network specifically consists of two 1D ConvNet layers followed by a fully connected layer. The performers-attention is then applied to effectively extract deep semantic correlations from action sequences involving human activities. After each layer, normalization and dropout regularization are applied to make the learning process faster and prevent the encoder from overfitting. 1D ConvNet-based networks have been proposed as fast and accurate models for HAR systems [11]. This is due to the ability of 1D ConvNet in extracting mostly correlated features by considering local dependency from temporal sequential input data.
2. Projection network $Proj(\cdot)$ maps the representation vector r to a projected vector $z = Proj(r) \in R^{D_E}$ where $D_E = 512$. The projector network is only a single fully connected layer appended to the encoder. The Encoder and projection networks are trained using contrastive loss function to make embeddings of similar classes are close together and dissimilar classes are far apart. The projection is discarded at the end of the contrastive training. Equation 3 shows the supervised contrastive loss function which is used in the first stage to learn the encoder.

$$\mathcal{L} = \frac{-1}{2N_{\tilde{y}_i} - 1}\sum_{j=1}^{2N}\mathbb{1}_{i\neq j}\cdot\mathbb{1}_{\tilde{y}_i=\tilde{y}_j} \cdot\log\frac{\exp(z_i\cdot z_j/\mathcal{T})}{\sum_{k=1}^{2N}\mathbb{1}_{i\neq k}\cdot\exp(z_i\cdot z_k/\mathcal{T})} \tag{3}$$

where

- N is the number of random samples in a mini-batch;
- N_y is the total number of samples in the mini-batch with the same label y;
- $z_i = Proj(E(x_i))$ and $z_j = Proj(E(x_j))$ are the projected vectors of the samples belonging to the same class;

- while $z_k = Proj(E(x_k))$ is the projected vector of a different class;
- $\mathcal{T}$ is a positive scalar temperature parameter;
- $\mathbb{1}_{i \neq j}$ avoids inner product of the same vector;
- $\mathbb{1}_{\tilde{y}_i = \tilde{y}_j}$ ensures that the z_i and z_j are the projected vectors of the same class;
- $\mathbb{1}_{i \neq k}$ is used to ensure that the z_k does not belong to the class of z_i and z_j.

In the second stage, a classifier with a fully connected layer followed by a softmax layer is trained using the encoder network. However, the encoder network of the first stage is frozen and the projector network is discarded. The learned representation from the encoder network without the projector network is used to learn the classifier. In the second stage, the network uses the focal loss function to predict human activities. The proposed network causal ConvNet based on supervised contrastive learning and Performer-attention forges recurrent settings to further accelerate the learning phase and improve recognition score for HAR systems. Causal convolution ensures that the model does not violate the ordering of the time steps of the temporal sensors data. The performers-attention supports the proposed network to pay extra attention to the discriminative features to accurately recognize human activities. Supervised contrastive learning is used to build the proposed network in two stages of learning, where the first stage is used to learn a good data representation for learning the classifier in the second stage. Two stages of learning are used to learn a better representation with more discriminative features that support the classifier to better distinguish human activities compared to a normal one stage learning. The focal loss function according to the effective number of examples is used to prevent skewed learning toward majority activities and improve the recognition scores of the minority activities.

5 Experiments and evaluation

In the section, experiments and evaluations based on eight datasets of human activities are shown and discussed. Moreover, results of the proposed network compared with the existing state-of-the-art models are shown.

5.1 Datasets and preprocessing

5.1.1 Ordonez smart environment datasets

Collected daily human activities in five intelligent environments using equipped sensors are used in this research to evaluate the proposed network. Ordóñez homes A and B [32] are two smart environments that are equipped with binary sensors to read and collect human activities. Different binary sensors within these two smart homes are utilized such as pressure sensors and passive infrared sensors to capture various human movements. The details of these two smart environments are shown in Table 1. In Ordóñez smart environment A, 12 binary sensors including PIR, pressure sensor, flush, and magnetic were employed to read and collect nine daily activities in 14 days over 20,358 minutes. In Ordóñez smart home B, ten human activities are recorded using 12 binary sensors in 22 days over 30,469 minutes. There are nine common activities from these two smart environments which are *Showering, Sleeping, Breakfast, Snack, Lunch, Spare Time/TV, Grooming, Toileting, and Leaving*. Besides, Ordóñez smart home B has one more recorded activity which is *Dinner*.

5.1.2 Kasteren smart environment datasets

Kasteren homes A, B and C are smart environments used to record human activities by embedded binary sensors [41]. The details of the recorded datasets from these smart

Table 1 Information about experimental datasets

	Ordonez home A -	Ordonez home B -	Kastern home A -	Kastern home B -	Kastern home C
Setting	Home	Home	Apartment	Apartment	House
Gender	–	–	Male	Male	Male
Activities	10	11	10	13	16
Age	-	-	26	28	57
Rooms	4	5	3	2	6
Sensors	12	12	14	23	21
Duration	14 days	21 days	25 days	14 days	19 days

Table 2 Details of the Ordonez smart home datasets

Human activity	Smart home A	Smart home B
Leaving	1,664	5,268
Snack	6	408
Grooming	98	427
Breakfast	120	309
Toileting	138	167
Showering	96	75
Idle	1,598	3,553
Lunch	315	395
Spare time/ TV	8,555	8,984
Dinner	-	120
Sleeping	7,866	10,763
Total	20,456	30,427

environments are shown in Table 1 regarding the activities, the number of sensors and residents. In Kasteren home A, ten human activities are recorded using fourteen binary sensors in 25 days over 40,005 minutes. In Kasteren home B, 13 human activities are captured using 23 binary sensors in 14 days over 38,900 minutes. In Kasteren home C, 16 human activities are captured using 21 binary sensors in 19 days over 25,486 minutes (Tables 2 and 3).

5.1.3 Wearable smartphone (inertial sensors) dataset

Inertial sensors are embedded in a waist-mounted smartphone to record the human activities of 30 participants [1, 35]. The age of participants is between 19 to 48 years old. The participants recorded six activities in which three activities are dynamic (walking downstairs, walking, and walking upstairs) and three activities are static postures (sitting, standing, lying). Samsung Galaxy S II as a wearable device is used by the participants to record their activities. To annotate the datasets the activities were video-recorded. 70% of participants' data are used for learning while 30% of participants' data are used for the inference phase. Table 4 shows the details of the training and testing sets for this datasets.

Table 4 Details of in the smartphone dataset

Human activity	Training samples	Testing samples
Walking_upstairs	1,073	471
Standing	1,374	532
Walking	1,226	496
Sitting	1,286	491
Laying	1,407	537
Walking_downstairs	986	420

5.1.4 Wearable wireless identification and sensing data

Human activities are recorded from 14 participants aged 78-82 years who wore Wearable Wireless Identification and Sensing Platform (W^2ISP) tag [39, 40]. Four activities which are i) sit on chair ; ii) ambulating ; iii) lying; iv) sit on bed are recorded. These activities are performed by senior people in two configured clinical rooms (*Roomset1* and *Roomset2*) that are used for ambulatory monitoring. The

Table 3 Details of human activities in the Kasteren smart homes

Activities	Home C	Activities	Home B	Activities	Home A
Get_dressed	70	Eat_brunch	132	Go_to_bed	11,599
prepare_dinner	300	Get_a_drink	6	Idle	7,888
Idle	5,883	Prepare_brunch	82	Get_snack	24
Prepare_breakfast	78	Prepare_dinner	87	Prepare_breakfast	59
Eating	345	Brush_teeth	25	Take_shower	221
Get_snack	8	Eat_dinner	46	Leave_house	19,693
Leave_house	11,915	Go_to_bed	6,050	Prepare_dinner	325
Prepare_lunch	58	Wash_dishes	25	Use_toilet	154
Go_to_bed	7,395	Idle	20,049	Brush_teeth	21
Take_shower	184	Leaving_the_house	12,223	Get_drink	21
Get_drink	20	Use_toilet	39		
Use_toilet_upstairs	35	Take_shower	109		
Take_medication	6	Get_dressed	27		
Shave	57				
Use_toilet_downstairs	57				
Brush_teeth	75				
Total	26,486	Total	38,900	Total	40,005

Table 5 Details of the wearable sensor datasets

Human activity	RoomSet1	RoomSet2
Ambulating	1,956	335
Sit on bed	15,162	1,244
Lying	30,983	20,537
Sit on chair	4,381	530

frequency distribution of activities from these two datasets (Roomset1 and Roomset) are shown in Table 5.

5.1.5 Preprocessing raw smart home sensors data

Recorded human activities from smart home environments are preprocessed where the timeline of activities are segmented with a window size Δt=1 minute. In the collected sensor data, sensor readings have start and end times. Moreover, the raw data also provides information about the type, location, and place of the sensors within the smart settings. To produce the input datasets from the collected sensor readings, a segmentation technique based on fuzzy temporal window (FTW) as a successful sliding window method is used [11, 13–15, 25, 34]. FTW as a data segmentation technique has been employed to extract sensor readings of short and long term performed activities such as preparing snacks or sleeping from collected sensors data [25, 34]. Temporal models have improved the performance of HAR systems when the FTW is used to generate model input datasets [11, 25].

5.2 Hyper-parameters of the proposed network

The proposed network uses these hyper-parameters, 128, 0.001 and 20% for the batch size, learning rate, and dropout rate, respectively to converge at the minimum of the validation loss. Early stopping as one of the techniques of regularization is used to determine the number of epochs and to prevent overfitting by stopping the training when the validation error of the proposed network starts increasing. The 20% dropout rate as another regularization technique after each learning layer is used to further avoid overfitting [38]. Batch normalization as a normalization technique is used to normalize the input data across the batches after each learning layer [18] to make deep learning models faster and more stable during training.

5.3 Evaluation of proposed network

To evaluate the proposed network against the existing state-of-the-art methods F1-score is used. Accuracy is a common metric to check the performance of the models, but accuracy is not a suitable metric to evaluate HAR systems due

Table 6 F1-score results in Ordonez home A dataset

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR + Attention	DCC+MSA	Proposed network
Breakfast	82.74	80.19	84.65	85.65	79.98	83.11	84.51	85.71	**86.89**
Grooming	46.66	57.14	51.28	74.21	62.19	75.32	80.00	80.01	**83.00**
Leaving	97.20	97.28	96.43	96.11	96.77	95.29	95.51	99.75	**99.81**
Lunch	95.65	96.92	94.87	95.42	95.34	95.44	94.39	96.93	**97.11**
Showering	78.94	75.45	77.94	79.42	78.12	80.65	86.89	93.84	**93.91**
Sleeping	96.77	96.34	96.63	95.57	94.89	97.53	97.11	97.63	**97.69**
Snack	64.66	67.22	55.83	67.02	69.99	70.74	82.63	84.82	**85.31**
Spare time	98.50	98.04	97.84	96.66	98.81	96.83	97.21	98.57	**98.82**
Toileting	63.75	61.09	64.71	62.71	67.42	69.89	77.25	79.76	**81.26**
Average	80.54	81.07	80.02	83.64	82.61	84.97	88.55	90.78	**91.53**

to the existence of imbalanced classes in human activities [11]. Therefore, F1-score is used to measure and evaluate the performance of the proposed network against the existing methods. F1-score is computed by recall with precision and provides a better measure of the incorrectly classified activities than the accuracy metric [25]. F1-score as a performance metric is used to evaluate the results of the experiments. The F1-score ($2\frac{\text{precision}\times\text{recall}}{\text{prscecision}+\text{recall}}$) is the weighted average of recall ($\frac{TP}{TP+FN}$) and precision ($\frac{TP}{TP+FP}$), where FN, FP, and TP are the number of false negatives, false positives and true positives, respectively. F1-score is commonly used in evaluating HAR systems [11, 13, 15, 25].

5.4 Results and discussion

The experimental results and findings of the proposed network are shown and discussed. The proposed causal ConvNet-based performers-attention and supervised contrastive learning for HAR are compared with several state-of-the-art methods: HAR+Attention [24], DeepConvLSTM+Attention [37], DCC+MSA [13] and many temporal models i.e. LSTM, 1D ConvNet, hybrid of 1D ConvNet and LSTM, Bi-LSTM, and CuDNN LSTM. The architectures and the results of temporal models are shown and reported in [13]. To evaluate the proposed causal ConvNet-based performers-attention and supervised contrastive learning against existing methods, eight benchmark human activity datasets are used. Tables 6 to 14 show that the results from all the datasets based on the proposed network outperform the existing methods. Moreover, the proposed network enhances the performance of the minority classes compared to the existing methods. The achieved results based on each of the datasets are separately discussed and evaluated in the following Sections.

To evaluate the proposed methods, the leave-one-day-out cross-validation is used for the smart home datasets as it is commonly used for HAR. The human activity recorded data for a single day are used to inference the model and the recorded data for the rest of the days are used to train the model. This technique is commonly used in HAR. Besides, K-fold cross-validation technique is used to evaluate the wearable sensors data since information about recording dates is not provided in the wearable sensors data. To show the results of the proposed model, the average F-score of the cross-validation is computed as done in the following research [11, 25, 33, 34].

5.4.1 Results from Ordóñez datasets

The outcomes of the experiments for the proposed network against the existing state-of-the-art methods based on the Ordóñez smart environments A and B are shown in Tables 6

Table 7 F1-score results in Ordonez home B dataset

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR + Attention	DCC+MSA	Proposed network
Leaving	88.34	93.05	90.87	89.52	92.86	89.79	92.09	93.33	**94.98**
Sleeping	98.11	85.71	86.76	83.29	86.82	96.37	95.42	98.30	**98.45**
Grooming	65.75	85.55	85.36	86.12	81.62	85.33	87.91	88.87	**90.11**
Breakfast	75.12	64.44	69.38	68.93	66.83	74.87	75.39	76.87	**78.46**
Showering	78.94	79.91	78.56	77.69	80.78	79.43	79.12	82.84	**83.32**
Lunch	98.95	81.18	77.00	79.68	83.21	95.21	95.31	99.63	**99.65**
Snack	67.92	75.86	73.41	75.42	73.21	76.16	76.31	78.59	**79.59**
Toileting	48.91	80.00	83.62	76.47	83.32	83.56	83.24	86.11	**88.74**
Spare time	74.63	78.51	77.24	73.32	77.98	78.21	79.32	81.48	**84.39**
Dinner	82.23	84.78	85.32	82.51	85.49	86.19	86.49	89.45	**90.22**
Average	77.89	80.89	80.75	79.29	81.21	84.51	85.06	87.34	**88.79**

Table 8 F1-score results of Kasteren smart home A dataset

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR + Attention	DCC+MSA	Proposed network
Get_Snack	50.00	53.21	51.23	55.71	56.42	57.22	58.71	63.69	**65.24**
Get_drink	51.76	56.84	48.87	42.81	57.21	59.33	59.54	66.92	**68.21**
Brush_teeth	20.22	24.08	37.86	21.56	31.46	43.59	52.22	54.44	**57.19**
Prepare_breakfast	76.66	74.51	72.41	74.95	75.57	76.97	79.54	83.32	**85.31**
Go_to_bed	79.72	74.63	73.21	78.8	73.31	80.16	81.76	86.54	**87.99**
Leave_house	79.80	81.58	80.28	76.37	78.89	80.02	82.19	84.28	**86.76**
Use_toilet	56.60	66.11	63.06	58.42	67.85	67.82	69.34	71.97	**73.21**
Take_shower	84.37	81.43	79.71	74.86	83.69	85.13	89.23	89.11	**90.87**
Prepare_dinner	83.20	85.24	80.39	87.94	86.48	89.56	91.87	95.42	**96.66**
Average	64.70	67.59	65.22	63.49	67.98	71.09	73.82	77.29	**79.04**

Table 9 F1-score results in Kasteren smart home B datasets

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR + Attention	DCC+MSA	Proposed network
Brush_teeth	23.10	37.62	33.25	32.57	39.55	42.89	47.82	51.18	**53.21**
Eat_brunch	88.42	90.14	87.53	91.72	89.87	90.93	91.11	95.92	**96.21**
Eat_dinner	83.19	85.23	86.68	86.01	86.31	86.79	86.29	90.02	**91.87**
Get_a_drink	17.84	31.18	22.34	25.61	33.03	44.15	44.75	53.00	**55.21**
Go_to_bed	95.11	99.01	99.21	98.91	97.94	96.32	94.48	99.73	**99.88**
Leaving_the_house	91.13	91.75	86.14	87.46	92.00	92.98	93.21	96.39	**97.78**
Prepare_brunch	77.48	80.19	83.11	85.92	79.96	85.62	84.29	88.10	**89.92**
Get_dressed	16.66	22.58	20.08	27.10	23.41	31.79	41.11	42.63	**45.89**
Prepare_dinner	93.11	97.29	94.90	97.00	96.87	96.21	95.31	97.51	**97.98**
Take_shower	76.82	79.12	82.71	75.91	78.91	81.95	82.13	83.13	**85.32**
Use_toilet	47.78	52.51	47.08	55.71	53.22	56.13	54.19	62.18	**64.86**
Wash_dishes	76.61	76.12	73.19	49.28	75.80	75.38	77.25	82.36	**84.29**
Average	65.63	70.22	68.01	67.76	70.57	73.42	74.32	78.51	**80.20**

Table 10 F1-score results in Kasteren home C datasets

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR + Attention	DCC+MSA	Proposed network
Eating	76.71	81.32	79.69	80.18	80.36	80.98	84.22	85.31	**86.89**
Brush_teeth	51.27	61.56	62.82	60.73	62.59	63.55	67.76	68.11	**69.98**
Get_dressed	53.47	55.90	51.47	54.78	56.32	56.82	60.27	61.17	**64.19**
Get_drink	42.13	47.61	50.40	38.99	47.91	48.11	50.37	51.71	**55.25**
Get_snack	64.14	67.74	65.53	68.16	68.39	67.86	70.45	72.23	**74.22**
Go_to_bed	94.86	95.11	91.48	94.21	96.04	95.41	93.21	96.12	**97.59**
Leave_house	93.81	90.18	92.74	89.05	91.52	92.57	91.39	94.17	**95.76**
Prepare_breakfast	76.35	75.74	73.15	77.78	76.81	78.45	81.24	83.42	**84.68**
Prepare_dinner	77.01	79.74	68.32	71.53	78.49	79.68	80.14	84.29	**86.21**
prepare_lunch	74.12	76.21	77.21	73.55	77.07	78.39	79.31	85.73	**87.63**
Use_toilet_downstairs	42.68	40.90	41.46	37.98	41.69	45.17	49.55	59.29	**62.59**
Use_toilet_upstairs	35.21	43.27	30.97	45.57	45.32	46.21	48.64	51.19	**53.44**
Shave	73.83	75.32	77.15	71.73	76.42	78.25	77.82	81.01	**83.41**
Take_medication	48.37	43.74	45.32	49.32	42.39	45.21	56.52	57.09	**61.36**
Take_shower	72.18	75.29	74.34	75.87	75.71	75.42	76.61	78.16	**80.99**
Average	65.02	67.30	65.47	65.96	67.80	68.79	71.16	73.93	**76.27**

Table 11 F1-score results in smartphone dataset

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR+Attention	DCC+MSA	Proposed network
Laying	89.03	87.14	86.76	85.35	87.51	89.67	89.91	95.69	**96.79**
Sitting	84.32	82.29	81.22	82.53	81.98	86.45	90.25	95.94	**95.99**
Standing	88.38	86.71	87.42	86.32	86.43	88.92	88.56	92.77	**94.32**
Walking	75.90	80.17	78.87	75.64	81.03	80.89	81.11	84.89	**86.89**
Walking_downstairs	76.31	80.01	79.11	78.76	80.21	80.11	83.91	84.14	**86.38**
Walking_upstairs	96.44	95.42	94.63	95.89	96.05	96.93	97.23	100.00	**100.00**
Average	85.89	86.14	85.00	84,08	86.03	87.16	88.49	92.24	**93.39**

Table 12 F1-score results in wearable dataset of RoomSet1

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR+Attention	DCC+MSA	Proposed network
Ambulating	92.91	92.58	93.67	92.02	93.19	93.67	95.22	97.63	**98.01**
Lying	93.94	91.34	87.94	94.71	92.97	94.12	95.21	97.70	**97.92**
Sit_on_bed	94.91	94.74	95.89	94.64	95.94	95.31	96.25	99.90	**99.93**
Sit_on_chair	56.51	49.12	47.48	51.58	50.04	60.52	70.11	72.84	**74.41**
Average	84.56	81.94	81.24	83.23	83.03	85.90	89.19	92.02	**92.56**

Table 13 F1-score results in wearable dataset of RoomSet2

Activity	ConvNet	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	HAR+Attention	DCC+MSA	Proposed network
Ambulating	79.42	83.75	84.95	78.95	84.67	85.43	87.11	89.79	91.93
Lying	89.75	82.29	89.50	84.97	83.16	89.42	89.32	94.85	95.32
Sit_on_bed	94.74	94.66	96.75	95.49	95.21	96.21	97.52	99.79	99.95
Sit_on_chair	51.31	48.27	59.70	53.75	49.51	62.56	68.87	69.87	72.31
Average	78.80	77.24	82.72	78.24	78.13	83.40	85.70	88.57	**89.87**

Table 14 Ablation study results of the proposed network

Datasets	Without performer attention	Without two stage learning	Without causal	Without focal loss	Proposed network
Ordonez home A	87.64	86.73	88.29	88.42	91.53
Ordonez home B	85.36	85.03	86.14	85.32	88.79
Kastern home A	75.26	75.00	77.84	75.96	79.04
Kastern home B	77.43	76.21	78.67	77.43	80.20
Kastern home C	74.35	73.46	75.03	74.39	76.27
Smartphone dataset	89.42	87.11	90.93	91.71	93.39
Wearable RoomSet1	88.19	88.23	89.24	87.18	92.56
Wearable RoomSet2	85.42	85.85	87.32	86.43	89.87

and 7. The results demonstrate that our proposed network obtained better results compared with many temporal models (LSTM, 1D ConvNet, hybrid, Bi-LSTM, and CuDNN LSTM) in addition to several existing methods [13, 24, 37] for HAR. The proposed network improves the result scores of all the activities particularly the minority classes. The minority classes such as *Snack, Grooming, Toileting, Showering, Dinner, and Breakfast* as shown in Table 2 are well improved using our proposed network compared to the existing methods. The proposed network achieved better average results for all classes in addition to the results of each activity in both of the smart home datasets.

5.4.2 Results from kasteren datasets

The results of the proposed network based on the datasets A, B, and C from Kasteren smart homes against the temporal models (LSTM, 1D ConvNet, hybrid 1D ConvNet + LSTM, CudNN LSTM and Bidirectional LSTM) in addition to the existing methods are shown in Tables 8, 9 and 10. The proposed network enhances the performances of each human activity and the average result score of all activities including the minority classes such as *Get_dressed, Get_snack* as shown in Table 3 compared with the existing methods.

5.4.3 Results from wearable sensors datasets

The results of the proposed network for HAR from wearable sensors data are compared with the results of the existing methods. Tables 11, 12 and 13 show the detailed results of our proposed network compared with the existing methods. The results of the proposed network from smartphone sensors data are shown in Table 11. The results of the wearable sensors data from Roomset1 and Roomset2 are shown in Tables 12 and 13 and demonstrate that the proposed network outperformed the state-of-the-art techniques. The proposed network enhanced the performance of the individual activity and the average performance of all activities compared to the existing methods from all wearable sensor data. Moreover, the proposed network improved the results of the minority class such *Sit on chair, Ambulating, and Walking_downstairs* compared with the existing methods.

5.4.4 Ablation study of the proposed network

An ablation study is completed to show the contribution of each component in the proposed network for HAR systems. The proposed network without performer attention, two stages learning, causal convolution, and focal loss. Table 14 demonstrate the results of the proposed network without these four components and the proposed network from the

Table 15 Training time in seconds of the proposed network based on self-attention and performer attention compared to existing methods

Activity	Proposed network		DCC+MSA	DeepConvLSTM+Attention	HAR+Attention
	Self-attention	Performer attention			
Ordonez home A	165.19	**131.56**	148.26	1012.42	1921.43
Ordonez home B	301.46	**242.76**	271.89	1241.42	2317.71
Kastern home A	156.75	**123.61**	148.05	1056.29	1873.56
Kastern home B	149.32	**119.21**	137.35	902.14	1564.12
Kastern home C	102.57	**73.19**	95.14	789.35	1373.89
Smartphone dataset	128.98	**101.14**	121.13	697.89	1125.89
Wearable RoomSet1	217.39	**169.37**	198.24	618.64	1078.32
Wearable RoomSet2	79.11	**49.87**	56.05	135.31	892.32

experimental datasets. The results indicate the impact of each component in the proposed network. For instance, the proposed network obtained the F1-score of 91.53, while the proposed network without performer attention obtained the F1-score of 87.64, without two stages learning obtained the F1-score of 86.73, without causal convolution obtained the F1-score of 88.29 and without using the focal loss, the F1-score is 88.42. This example confirms the contribution of supervised contrastive learning. Moreover, the proposed network without using two stages of learning has gained the lowest results from the sensor datasets compared with other components of the proposed network. Hence, the results show that the higher contribution is made by the proposed supervised contrastive learning with two stages of learning in the proposed network compared to the performers-attention, causal convolutions, and focal loss.

5.4.5 Learning time of the proposed network

The training time of our proposed network to converge with the smallest validation loss based on the self-attention and performers-attention is reported. The learning time of our proposed network is compared with DeepConvLSTM+Attention [37], DCC+MSA [13], and HAR+Attention [24] methods. The results of the experiments show that the learning time of the proposed network based on the performers-attention is lower than the training time of the proposed network based on the self-attention. In addition, the learning time of the proposed network to converge is also lower than the learning time of the existing methods as shown in Table 15. For example the proposed network based on the performers-attention converged in 131.56 seconds while our proposed network based on the self-attention is converged in 165.19 seconds. Therefore, our proposed network is faster than the methods proposed based on the self-attention mechanism.

6 Conclusion

This study proposes causal ConvNet-based performers-attention and supervised contrastive learning to improve human activity recognition and reduce the training time in the datasets collected from smart home environments and wearable sensors. Extensive experiments are performed on eight datasets to evaluate the proposed network compared to the basic temporal models and existing state-of-the-art methods. The proposed network has four main components which are: causal convolution, performers-attention, supervised contrastive learning for two stages of learning (representation learning and classifier learning), and focal loss. Causal convolution is used to preserve the ordering

of the input temporal data which is significant for human activity recognition. The performers-attention is used in the proposed network to focus more on the important timesteps to improve the recognition process. Supervised contrastive learning is used to prepare a discriminative representation and further reduce the classification error compared with several existing methods for human activity recognition. Further, the focal loss function is used to address imbalanced activities problems and improve the less presented human activities. The results of the thorough experiments reveal that the proposed network outperforms the current methods and reduced the learning time compared with the existing state-of-the-art methods. We further performed ablation studies to highlight the contribution of each component of the proposed network. The results of the ablation studies show that the proposed supervised contrastive learning with two stages of learning provides a larger contribution in our proposed network compared with the performers-attention, causal convolutions, and focal loss.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL (2013) A public domain dataset for human activity recognition using smartphones. In: Esann, vol 3, p 3
2. Bai S, Zico Kolter J, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271
3. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
4. Betancourt C, Chen W-H, Kuan C-w (2020) Self-attention networks for human activity recognition using wearable devices. In: 2020 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 1194–1199
5. Cao L, Wang Y, Bo Z, Jin Q, Vasilakos AV (2018) Gchar: an efficient group-based context—aware human activity recognition on smartphone. J Parallel Distrib Comput 118:67–80
6. Choromanski K, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, Hawkins P, Davis J, Mohiuddin A, Kaiser L et al (2020) Rethinking attention with performers. arXiv:2009.14794
7. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9268–9277
8. Fang H, Si H, Chen L (2013) Recurrent neural network for human activity recognition in smart home. In: Proceedings of 2013 Chinese intelligent automation conference. Springer, pp 341–348
9. Gao W, Zhang L, Teng Q, Wu H, Min F, He J (2020) Danhar: dual attention network for multimodal human activity recognition using wearable sensors. arXiv:2006.14435
10. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 2. IEEE, pp 1735–1742
11. Hamad RA, Salguero AG, Bouguelia M, Espinilla M, Quero JM (2019) Efficient activity recognition in smart homes using delayed fuzzy temporal windows on binary sensors. IEEE J Biomed Health Inf: 1–1. https://doi.org/10.1109/JBHI.2019.2918412
12. Hamad R, Jarpe E, Lundstrom J (2018) Stability analysis of the t-sne algorithm for human activity pattern data. In: 2018 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 1839–1845
13. Hamad RA, Kimura M, Yang L, Woo WL, Bo W (2021) Dilated causal convolution with multi-head self attention for sensor human activity recognition. Neural Comput Appl 33(20):13705–13722
14. Hamad RA, Kimura M, Lundström J (2020) Efficacy of imbalanced data handling methods on deep learning for smart homes environments. SN Comput Sci 1(4):1–10
15. Hamad RA, Yang L, Woo WL, Wei B (2020) Joint learning of temporal models to handle imbalanced data for human activity recognition. Appl Sci 10(15):5293
16. He J, Zhang Q, Wang L, Pei L (2018) Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. IEEE Sensor J 19(6):2287–2297
17. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
18. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp 448–456. PMLR
19. Jiang W, Yin Z (2015) Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of the 23rd ACM international conference on multimedia, pp 1307–1310
20. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D (2020) Supervised contrastive learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Inc., pp 18661–18673. https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf
21. Lee D, Helal S (2013) From activity recognition to situation recognition. In: International conference on smart homes and health telematics. Springer, pp 245–251
22. Liciotti D, Bernardini M, Romeo L, Frontoni E (2020) A sequential deep learning application for recognising human activities in smart homes. Neurocomputing 396:501–513
23. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
24. Mahmud S, Tonmoy MTH, Bhaumik KK, Rahman AM, Amin MA, Shoyaib M, Khan MAH, Ali A (2020) Human activity recognition from wearable sensor data using self-attention. In: ECAI 2020 - 24th European conference on artificial intelligence, 29 August-8 September 2020. Santiago de Compostela, Spain

25. Medina-Quero J, Zhang S, Nugent C, Espinilla M (2018) Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition. Expert Syst Appl 114:441–453
26. Murahari VS, Plötz T (2018) On attention models for human activity recognition. In: Proceedings of the 2018 ACM international symposium on wearable computers, pp 100–103
27. Niu W, Long J, Han D, Wang Y-F (2004) Human activity detection and recognition for video surveillance. In: 2004 IEEE international conference on multimedia and expo (ICME)(IEEE cat. no. 04TH8763), vol 1. IEEE, pp 719–722
28. Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. Springer, pp 69–84
29. Nweke HF, Teh YW, Al-Garadi MA, Alo UR (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl 105:233–261
30. Ogbuabor G, La R (2018) Human activity recognition for healthcare using smartphones. In: Proceedings of the 2018 10th international conference on machine learning and computing, pp 41–46
31. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N (2016) Andrew senior, and koray kavukcuoglu
32. Ordóñez F, De Toledo P, Sanchis A et al (2013) Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. Sensors 13(5):5460–5477
33. Ordóñez FJ, Roggen D (2016) Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1):115
34. Quero JM, Orr C, Zang S, Nugent C, Salguero A, Espinilla M (2018) Real-time recognition of interleaved activities based on ensemble classifier of long short-term memory with fuzzy temporal windows. In: Multidisciplinary digital publishing institute proceedings, vol 2, p 1225
35. Reyes-Ortiz J-L, Oneto L, Samà A, Parra X, Anguita D (2016) Transition-aware human activity recognition using smartphones. Neurocomputing 171:754–767
36. Singh D, Merdivan E, Hanke S, Kropf J, Geist M, Holzinger A (2017) Convolutional and recurrent neural networks for activity recognition in smart environment. In: Towards integrative machine learning and knowledge extraction. Springer, pp 194–205
37. Singh SP, Sharma MK, Lay-Ekuakille A, Gangwar D, Gupta S (2020) Deep convlstm with self-attention for human activity decoding using wearable sensors. IEEE Sensor J 21(6):8575–8582
38. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
39. Torres RLS, Ranasinghe DC, Shi Q, Sample AP (2013a) Sensor enabled wearable rfid technology for mitigating the risk of falls near beds. In: 2013 IEEE international conference on RFID (RFID). IEEE, pp 191–198
40. Torres RLS, Ranasinghe DC, Shi Q (2013b) Evaluation of wearable sensor tag data segmentation approaches for real time activity classification in elderly. In: International conference on mobile and ubiquitous systems: computing, networking, and services. Springer, pp 384–395
41. van Kasteren TLM, Englebienne G, Kröse BJA (2011) Data: Human activity recognition from wireless sensor network Benchmark and software. In: Activity recognition in pervasive intelligent environments. Springer, pp 165–186
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv:1706.03762

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rebeen Ali Hamad received the BSc. degree in computer science from the University of Sulaimani in 2011. In 2016 he obtained an MSc in Computer Science with Distinction at The University of Nottingham, he was able to do this through a prestigious Higher Committee for Education Development in Iraq (HCEDiraq) scholarship, which is only awarded to talented and highly-ranked Iraqi students to support them in higher education. In June 2017, he joined the Center for Applied Intelligent Systems Research (CAISR) at Halmstad University. He worked on the Situation Awareness for Ambient Assisted Living (SA3L) project, which aims to perform robust recognition of dangerous situations and detect deviations of behaviour to enhance elderly-care alert systems. He is currently pursuing a PhD degree in computer and information science at Northumbria University. His research interests include knowledge representation, transfer learning, unsupervised learning, and activity recognition.

Longzhi Yang (Senior Member, IEEE) received the B.Sc. degree from the Nanjing University of Science and Technology, Nanjing, China, the M.Sc. degree from Coventry University, Coventry, U.K., and the PhD degree from the University of Wales, Aberystwyth, U.K., all in computer science in 2003, 2006, and 2011, respectively. He is a professor in the Department of Computer and Information Sciences at Northumbria University, Newcastle, U.K. His research interests include computational intelligence, machine learning, big data, cybersecurity, computer vision, intelligent control systems, robotics and the application of such techniques in real-world uncertain environments. He is a Senior Fellow of the Higher Education Academy and the Founding Chair of the IEEE Special Interest Group on Big Data for Cyber Security and Privacy.

Wai Lok Woo (Senior Member, IEEE) received the B.Eng. degree in electrical and electronics engineering and the M.Sc. and Ph.D. degrees in machine learning from Newcastle University,U.K., in 1993, 1995, and 1998, respectively. He is currently a Professor of Machine Learning with Northumbria University, U.K., where he leads the Research Cluster of Artificial Intelligence and Digital Technology. His major research interest includes the mathematical theory and algorithms for data science and analytics which includes areas of artificial intelligence, machine learning, data mining, latent component analysis, and multidimensional signal and image processing. He is also a member of the Institution of Engineering Technology. He was a recipient of the IEEE Prize and the British Commonwealth Scholarship. He serves as an Associate Editor to several international journals, including Sensors, IET Signal Processing, the Journal of Computers, and the Journal of Electrical and Computer Engineering.

Bo Wei has been a lecturer (assistant professor) in the School of Computing and Communications at Lancaster University, UK. He was a postdoctoral research assistant at University of Oxford and a senior lecturer (assistant professor) at Northumbria University. He obtained his PhD degree in Computer Science and Engineering in 2015 from the University of New South Wales, Australia. His research interests are mobile computing, Internet of Things, cyber security and wireless sensor networks.