Neural Computing and Applications (2021) 33:13705–13722
https://doi.org/10.1007/s00521-021-06007-5

ORIGINAL ARTICLE

Dilated causal convolution with multi-head self attention for sensor human activity recognition

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

Received: 10 December 2020 / Accepted: 31 March 2021 / Published online: 19 April 2021

Abstract

Systems of sensor human activity recognition are becoming increasingly popular in diverse fields such as healthcare and security. Yet, developing such systems poses inherent challenges due to the variations and complexity of human behaviors during the performance of physical activities. Recurrent neural networks, particularly long short-term memory have achieved promising results on numerous sequential learning problems, including sensor human activity recognition. However, parallelization is inhibited in recurrent networks due to sequential operation and computation that lead to slow training, occupying more memory and hard convergence. One-dimensional convolutional neural network processes input temporal sequential batches independently that lead to effectively executed operations in parallel. Despite that, a one-dimensional Convolutional Neural Network is not sensitive to the order of the time steps which is crucial for accurate and robust systems of sensor human activity recognition. To address this problem, we propose a network architecture based on dilated causal convolution and multi-head self-attention mechanisms that entirely dispense recurrent architectures to make efficient computation and maintain the ordering of the time steps. The proposed method is evaluated for human activities using smart home binary sensors data and wearable sensor data. Results of conducted extensive experiments on eight public and benchmark HAR data sets show that the proposed network outperforms the state-of-the-art models based on recurrent settings and temporal models.

Keywords Activity recognition · Smart home · Self-attention · Dilated causal convolution

1 Introduction

Human activity recognition (HAR) is a significant research field in ubiquitous computing for monitoring behaviors of people which plays an important role in various applications such as healthcare monitoring [1], security surveillance system [2] and resident situation assessment [3]. In healthcare monitoring, HAR, as one of the significant applications of intelligent environment and wearable sensor technologies, has been used to monitor the activity of daily living (ADL) in order to support and assist senior people, disabled and cognitive impaired [4]. HAR from smart home setting equipped by ubiquitous sensors in the field of ambient assisted living has gained increased attention for improving the quality of independent living of the residents within the smart home environment [5]. Smart homes with unobtrusive sensing technology for HAR have been used as a suitable solution for enhancing independent living when privacy is concerned [4, 6]. Wearable sensors

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

Masashi Kimura
kimura@convergence-lab.com

Longzhi Yang
longzhi.yang@northumbria.ac.uk

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

Bo Wei
bo.wei@northumbria.ac.uk

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

[2] Convergence Lab. Inc., Tokyo, Japan

are mainly embedded into mobile devices, wristwatches, clothes, glasses, belts, or shoes. Wearable sensors can be worn on the human body to capture their interaction with their physical surroundings, motion, posture, and location. Wearable sensors such as gyroscope, accelerometer, GPS, and RFID-readers (used together with RFID tags) have commonly been used to record information about users' movement e.g., walking, running, and laying [7]. Wearable sensors have been used for HAR since they can collect information such as body movement and position [7, 8]. The aim of HAR is to identify and recognize simple and complex human daily activities using smart home and wearable sensors data. HAR based on the sensors' data is a challenging task since sometimes the data could be noisy which leads to ambiguity in the interpretation of human activities [9]. Noise in the data could be caused by errors in the sensor connection system which fails to provide correct sensor activations. HAR systems based on sensors' data have been notably progressed and obtained promising results by the current development of machine learning in elderly-care alert systems and assistance in emergencies [10]. Monitoring long-term daily routine activities of a resident in a smart home setting provide utility to determine and assess wellness. Particularly, remote monitoring of daily routine activities such as eating, sleeping, or medication intake enables caregivers to track and assess the functional health status and to support the needs of the elderly people living alone [9]. Moreover, smart home and wearable sensors are able to provide sufficient information to properly detect the postural and ambulatory activities [11, 12].

Traditional machine learning approaches such as naive Bayes, support vector machine, and hidden Markov models have made tremendous progress on HAR and obtained satisfying results [8]. However, these approaches entirely rely on hand-crafted heuristic feature extraction, which is highly data dependent, usually limited by domain experts. Handcrafted features are not always generalizable across application domains and time consuming. Moreover, handcrafting does often not generate a sufficient number of features from a given dataset [13]. Recently, deep learning methods have been increasingly used in various applications of computer vision [14] and natural language processing, speech [15] and audio recognition [16]. Besides, deep learning methods have been used for HAR systems based on smart home sensors and wearable sensors. Most of these HAR systems have shown encouraging results for different purposes and from different datasets of daily routine activities. Among the deep learning methods, long short-term memory (LSTM) as a sequential deep learning model and variation of recurrent neural network (RNN) has achieved state-of-the-art performance for temporal information processing in various applications [17]. Particularly, LSTM for recognizing activities of daily living (ADLs) shows state-of-the-art performance on various activity recognition benchmark datasets using wearable sensors and smart home sensors [6, 13]. ADLs are introduced as the normal daily activities where we perform for self-care such as eating, drinking, and bathing [18]. Even though LSTM models improve the performance of HAR systems, training LSTM models are computationally expensive due to using the gating mechanism that allows long-term dependencies. Moreover, LSTM models occupy more memory and cannot process timesteps of input data in parallel since each timestep needs the results of the previous timestep to be processed [19, 20]. One dimensional convolutional neural network (1D CNN) has been used instead of LSTM to capture the sequential temporal information in the input data for HAR systems [21, 22]. Despite training of 1D CNN models are extremely faster compared to recurrent methods such as LSTM due to the absence of recurrent connections, the achieved results based on 1D CNN fall short of the results shown by LSTM in HAR systems. Moreover, 1D CNN is not sensitive to the timestep order which is the key in HAR systems. To address these problems, we propose dilated causal self-attention convolution that entirely forgoes recurrent settings to improve the performance of HAR. We adopted dilated causal convolution which is used as a part of the WaveNet to generates raw audio waveforms [23]. Dilated causal convolution is used to allow long-range temporal dependencies in the WaveNet that outperforms LSTM [24]. While dilated causal convolution captures long-range temporal sequential information [25], it is crucial to focus on particular information from the feature maps generated by dilated causal convolution using the self-attention mechanism [26]. The self-attention mechanism that is leveraged by transformer [19] can enable temporal models to expose context from the feature maps within the sequence.

To summarise, the main contributions of this paper are:

i. Proposing a model to accelerate training time and improve the results of activity recognition compared to state-of-the-arts using dilated causal and self-attention convolution.
ii. Causal convolutions within the proposed method are used to prevent information leakage from future to past.
iii. Dilated convolutions within the proposed method are used to maximize the receptive field by orders of magnitude and aggregate multi-scale contextual information without considerably increasing computational cost.
iv. Multi-head self-attention within the proposed method is used to effectively expose deep semantic

correlations from action sequences involving human activities.

v. Conducting extensive experiments using eight benchmark datasets of human daily activities from smart homes and wearable sensors to validate the proposed approach, which shows our proposed method can improve the accuracy by 5% up to 9% and reduce the training time compared with recurrent neural network-based architecture methods.

2 Related work

Human action recognition is a challenging research area based on sensor data and has attracted much attention in machine learning fields. Numerous methods have been proposed to model and recognize ADLs [8, 27]. Early research modeled activity recognition using support vector machine(SVM), decision tree, k-nearest neighbor (KNN) naïve Bayes [28]. HAR systems based on these traditional approaches have gained reasonable recognition results. However, these approaches solely process extracted heuristic-manual features of human activities. Hand-crafted features are usually limited by the availability of knowledge domain experts and a time-consuming task. Hence, deep learning models have been proposed in various applications to address these problems [29].

Deep learning models have shown satisfying results and reported state-of-the-art accuracy obtained on various HAR benchmark datasets [4, 6]. Moreover, deep learning models have been jointly used to handle imbalanced data and improve generalization of HAR [18]. Recurrent network-based architectures such as RNN and LSTM have been firmly established as state-of-the-art methods in sequence problems modeling including activity recognition [19]. RNN is employed to recognize the human daily activities from smart home sensor data [30]. The results show that RNN is useful in modeling and recognizing human activities. Yet RNN cannot properly process very long sequences and suffers from both gradient vanishing and exploding problems [4]. LSTM that solves vanishing and exploding gradient by the capability of handling long-term dependencies is often used to process temporal sequential data [31]. For instance, satisfying results are shown by employing LSTM to recognize human activities on diverse collected sensor data [4, 6, 32, 33]. Further, different LSTM architectures are proposed to improve the performance of HAR systems such as stacked LSTM [34], bidirectional LSTM [33], ensemble LSTM [35]. Moreover, combined CNN with LSTM is also employed to further improve the performance of HAR systems [6, 28].

The Dilated CNN can be used instead of the standard CNN to increase the convolution receptive field without losing resolutions [27]. Since the dilated convolution only appends empty elements between the elements of the standard convolution kernel, extra computational cost is not required for dilated convolution process. Dilated convolution is proposed [36, 37] for human activity recognition from wearable sensors data. Dilated convolution is also used for voice activity detection and audio source separation [38, 39].

Weakly supervised learning based on combined CNN and LSTM with self-attention layers using reinforcement learning trained on wearable sensors data for human activity recognition form [40]. However, this method requires large computing resources and has high complexity because it works based on reinforcement learning. Moreover, Convolution LSTM with self-attention mechanism is proposed to capture the spatio-temporal context in human activities and to focus on significant timesteps from temporal wearable sensors data [13]. References [41, 42] employed a self-attention mechanism to improve the performance of HAR systems based on wearable sensors. The results are improved using self-attention compared to the state-of-the-art. Betancourt et al. proposed an LSTM model based on the attention mechanism. The model is only tested on two wearable sensor datasets. The limitation of these methods is firstly the recurrent setting from the proposed methods leads to slow down the learning process. Secondly, the proposed methods from these studies are only applied to HAR systems based on wearable sensors, and the learning time is not considered. Reference [43] proposed DeepConvLSTM model based on attention mechanism for human activity recognition on smart home datasets. The method considered each time-step in the sequential data as a word and a specified time-window as the sentence. However, the method is evaluated only on three smart home datasets and only compared to bidirectional LSTM. Moreover, due to the sequential operation of recurrent settings in this method, parallelization is limited which makes this model computationally expensive. Gao et al. proposed a CNN dual attention without a recurrent setting for HAR systems, however, the method is only tested on wearable sensor datasets [44].

To address these limitations of HAR and improve the performance as well as reduce the learning time of HAR systems from smart homes sensor and wearable sensor data, we propose dilated causal and multi-head self-attention convolution.

3 Background

In section, we describe the temporal models i.e., 1D CNN, LSTM, and the hybrid 1D CNN and LSTM model.

3.1 Temporal modeling via LSTM

LSTM is an artificial RNN and used to learn from temporal sequential data. LSTM can handle and learn from long-term dependencies which alleviate vanishing and exploding gradient problem [31]. LSTM as a temporal model has been used to recognize ADLs from sensor data [4, 6]. LSTM processes temporal data using forget gate, input gate, and output gate to append or delete information to the cell state throughout the processing of the sequence data. The cell state is the main part of LSTMs that carry and transfer relevant information from earlier timesteps to later timesteps. Figure 1 shows the connection of the gates with the cell state in a single LSTM cell. The gates learn to keep relevant information and forget irrelevant information during training to update the information on the cell state. Hence each LSTM cell works as a memory to remove, read, and write information that is controlled by the forget, output, and input gates, respectively. Forget gate process both inputs the previous output h_{t-1} and new time step X_t using sigmoid activation function to indicate relevant or irrelevant information. The forget gate keeps the information if the outcome of the sigmoid function is 1 while deletes the information if the outcome of the sigmoid function is 0. Equation (1) shows how the forget gates within a single LSTM cell is computed. The next step consists of two parts to determine new information kept in the cell state. The first part is the input gate that indicates new information from the current input (X_t, h_{t-1}) is appended to the cell state. The tanh activation function is the second part that renders $\tilde{C}_t$ a vector of new candidate values and can be added to the cell state. Equations (2) and (3) show how the input gate and the new candidate values are computed, respectively. A new cell state C_t is generated based on the summation of the multiplication of these two parts and the multiplication of the forget gate with the previous cell state C_{t-1}. Equation (4) shows how the new cell state is computed. The multiplication of the previous cell state with the forget gate deletes part of the information which was decided to be forgotten earlier. Then the new candidate values are scaled by how much the cell state is updated using $it \times \tilde{C}_t$. Finally, the sigmoid activation function processes both the previous hidden state h_{t-1} and the current input timestep x_t to produce the output gate.

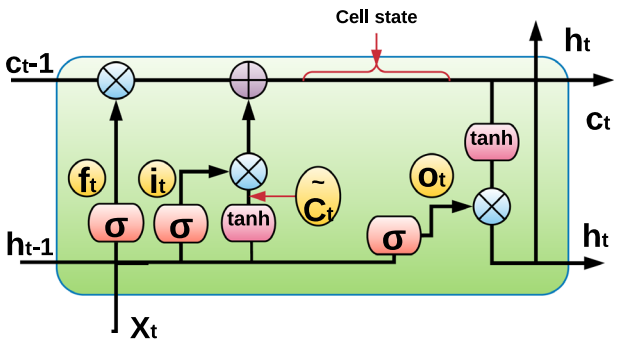

Fig. 1 Single LSTM cell

Finally, the output gate is computed based on filtered information using two different activation functions and also specifies the next hidden state. Then the tanh activation function processes the newly updated cell state. The output of the tanh functions multiplies by the output of the sigmoid function to render the next hidden state. The updated cell state and the newly generated hidden state pass information to the next timestep. Equations (5) and (6) show how the calculation of output gate and hidden state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \times \tanh C_t \tag{6}$$

where x is the input data, σ is the sigmoid activation function, tanh is the hyperbolic tangent activation function, W is the weight matrix.

LSTM has been used for HAR application and achieved promising results [4, 6, 9, 45]. Hence in this paper, LSTM as a temporal model is used to be compared with the proposed method. Two layers of LSTM with a flattened layer are stacked. Then the outputs of the flattened layer are passed into a fully connected layer with ReLU activation function and followed by a softmax layer. Figure 2 shows the architecture of the LSTM model.

Fast LSTM implementation backed by cuDNN (CUdNNLSTM) [46] is also used in this study with the same architecture of LSTM model. CUdNNLSTM is a version of LSTM that uses the CuDNN library, and it can only be run on a GPU to accelerate training and inference time.

3.2 Temporal modeling via 1D CNN

1D CNN has been widely used in HAR systems and has shown satisfying results [6, 21]. 1D CNN can properly extract features from raw and consider local dependency that is likely to be correlated. 1D CNN can also learn

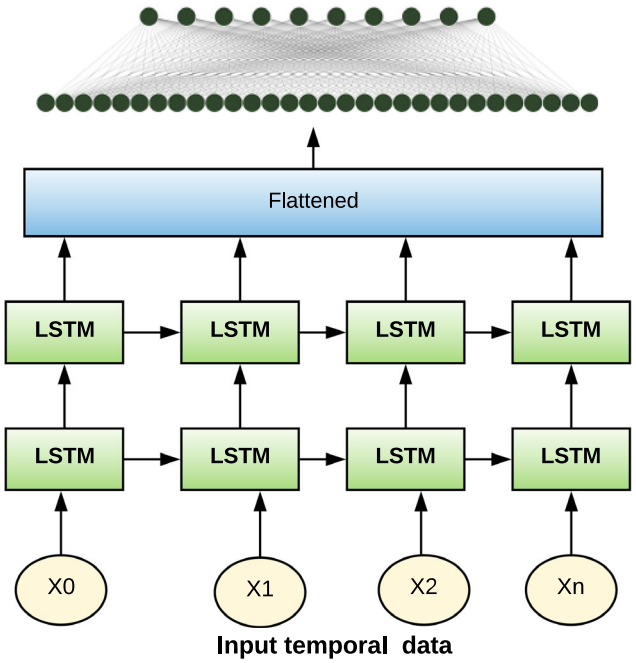

Fig. 2 Architecture of the LSTM model

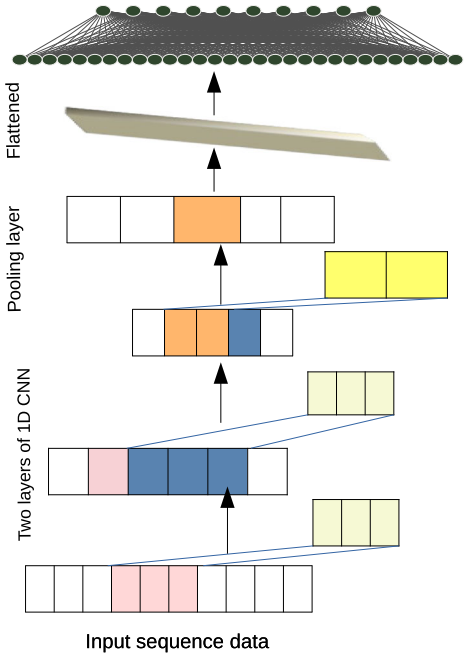

Fig. 3 Architecture of the 1D CNN model

hierarchical data representations of human activities that lead to improving HAR systems [45]. 1D CNN compared to LSTM has obtained competitive results in several applications such as activity recognition, machine translation, and audio generation with much faster learning time. However, 1D CNN is not sensitive to order that is significant in activity recognition [8]. Hence 1D CNN alone is not an optimal solution instead of LSTM. In this paper, 1D CNN is employed, and its results are shown. The 1D CNN model is designed by stacking two convolutional layers each with 64 filters. The kernel size of the 1D CNN in this study is equal to 3 that indicates the length of the 1D convolution window with stride size of 1. A Max-pooling layer with the window size equal to 2 is applied after the convolution layers to down-sample the features maps. The feature maps are flattened to be processed by the fully-connected, i.e., a dense layer with ReLU activation function followed by a soft-max layer. Figure 3 shows the architecture of 1D CNN.

3.3 Temporal modeling via Hybrid: 1D CNN + LSTM

The hybrid model based on stacking 1D CNN and LSTM sequentially has been used to improve the performance of HAR system [6, 18]. In this study, the hybrid model is employed by stacking one layer of each 1D CNN and LSTM to human activities from smart home data. Figure 4 shows the architecture of the hybrid model. The input data are firstly fed into the 1D CNN layer to extract features

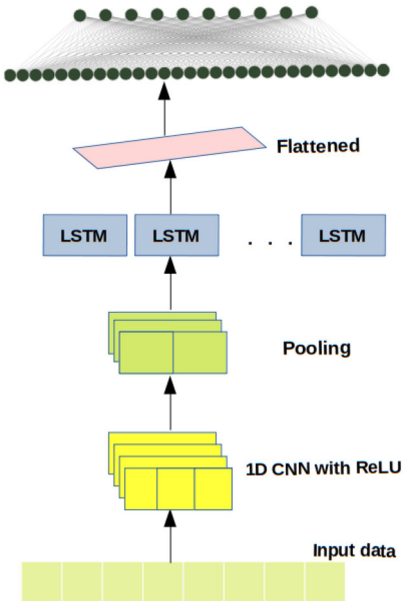

Fig. 4 Hybrid 1D CNN + LSTM model

before the LSTM layer to support sequence recognition. The input sub-sequences sensor data are processed independently by 1D CNN hence timestep orders are not considered. The feature maps of 1D CNN are down-sampled by a max-pooling layer with the window size equal to 2 before the LSTM layer. The feature maps are processed by the LSTM and then flattened followed by fully-connected layers, i.e., a dense layer with ReLU activation function and a soft-max layer. Furthermore, 1D CNN layers in the

hybrid model are often applied when recurrent-based models cannot realistically handle and process long-term dependencies from input sequence data. In such cases, 1D CNN in the hybrid model can make the long-term dependencies shorter through down-sampling by extracting higher-level features. Then the extracted features generated by 1D CNN could be better processed by the recurrent-based models [47]. However, order sensitivity is not considered in the extracted features by the 1D CNN. Hence, the hybrid of 1D CNN and LSTM is not the most acceptable solution to improve the performance of activity recognition [18].

3.4 Temporal modeling via Bidirectional LSTM

Bidirectional LSTM trains input data in forward and backward directions by using previous and subsequent information of a specific time step in two separate recurrent layers [48]. Figure 5 shows bidirectional LSTM where inputs of backward states are not connected to the outputs of the forward states. Including future information in addition to past information in bidirectional LSTM appears at first sight to violate causality [49]. Although Bidirectional LSTM has been successfully proposed in HAR and achieved satisfying results, Bidirectional LSTM is indeed expensive to train since it has a double recurrent setting in each layer [33]. Bidirectional LSTM is used in this study by stacking two forward and backward LSTMs layers. The outputs of these two layers are flattened and then fed to a fully-connected layer, i.e., a dense layer with ReLU activation function and a soft-max layer.

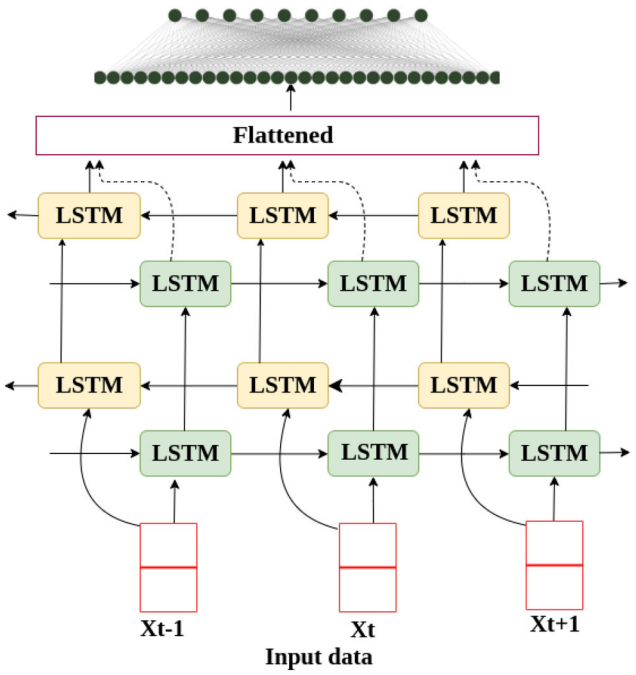

Fig. 5 Bidirectional LSTM model

4 Proposed method

In this section, we describe the proposed method, dilated causal convolution, and self-attention mechanism, for HAR in smart home data. We aim to design and propose a more efficient convolutional network model better than recurrent-based architecture models in terms of recognition score and training time. The distinguishing characteristics of our proposed method are: (1) the proposed model stops information leakage from future to past using causal convolution; (2) the proposed model can handle temporal sequential data of any length and map it to a series output of the same length; (3) the model can simultaneously focus on different important time steps of the sequence input using the multi-head self-attention mechanism. The details of the proposed model are described in the following subsections.

4.1 Sequence modeling

Before describing the details of the proposed model, we show the sequence modeling task for human activities. Input human activity sequences $x_0, \ldots, x_T$ are fed into a model to predict corresponding activity outputs $y_0, \ldots, y_T$ at each time. Predicting the activity output y_t for particular time t should be derived only by considering the observed times steps before time t: $x_0, \ldots, x_t$ [20]. Hence, sequence modeling is a function $f : x_0, \ldots, x_T \rightarrow y_0, \ldots, y_T$ (where x and y are the input and output, respectively) that renders the mapping as shown in Eq. (7).

$$\hat{y}_0, \ldots, \hat{y}_T = f(x_0, \ldots, x_T) \tag{7}$$

The model f is expected to minimize a loss L between, $L(\hat{y}_0, \ldots, \hat{y}_T, f(x_0, \ldots, x_T))$, the actual label and the predicted outputs where the input sequential data and the outputs are rendered based on some distribution. This formalism could not directly be used for domains such as sequence-to-sequence prediction or machine translation since these domains require the entire sequence input (past and future states) [20]. However, the setting can be extended for these domains.

4.2 Dilated causal convolutions

Causal convolutions used in the proposed method to control the model and predict output at time t based on only the convolutions of the sequence inputs from time t and earlier in the previous layers [20]. Causal convolutions also preserve the ordering of sequential input patterns. However,

causal convolutions require very large filters or many hidden layers to expand the receptive field [23]. To maximize the receptive field and aggregate multi-scale contextual information without considerably increasing computational cost, dilated convolutions are integrated into the proposed method. Dilated convolutions enable the model to increase the receptive field exponentially using a few layers and keeping the computational efficiency [25]. The dilated causal convolution *DCC* for one dimensional input sequence $x \in R^n$ with a filter $f : \{0, \ldots, k-1\} \rightarrow R$ on element s of the sequence is defined as:

$$DCC(x \bigstar_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i} \tag{8}$$

where d is the dilation factor, k is the filter size, and $s - d \cdot i$ shows the past direction. The dilation factor d is exponentially increased when the depth of the model is increased i.e., $d = 2^l$ at layer l of the model. Formally, we increase the dilation factors d exponentially by a factor of 2 in each layer $l = 1, \ldots, L$ where L is the number of layers of the dilated causal convolutions in the proposed model. Equation (9) shows the dilation factor in this study.

$$d \in [2^0, 2^1, 2^2, \ldots, 2^{L-1}] \tag{9}$$

In addition, the dilation convolution renders the standard convolution when $d = 1$. Figure 6 shows the dilation causal convolutions in the proposed model for dilations 1, 2, and 4. Dilated convolution with different dilation factors can be integrated with a filter at different ranges. The filters convolve input values over an area larger than its length using dilated convolutions by skipping input values with a certain step which is the dilation factor. Hence, dilation convolution is equivalent to a standard convolution with one dilating, but importantly more efficient. Dilation convolution effectively enables the model to aggregate multi-scale contextual information with fewer layers and the same receptive field compared to a standard convolution [25]. Therefore, the number of learnable parameters is reduced by using stacked dilated causal convolutions that lead to yield more efficient training and light-weight model.

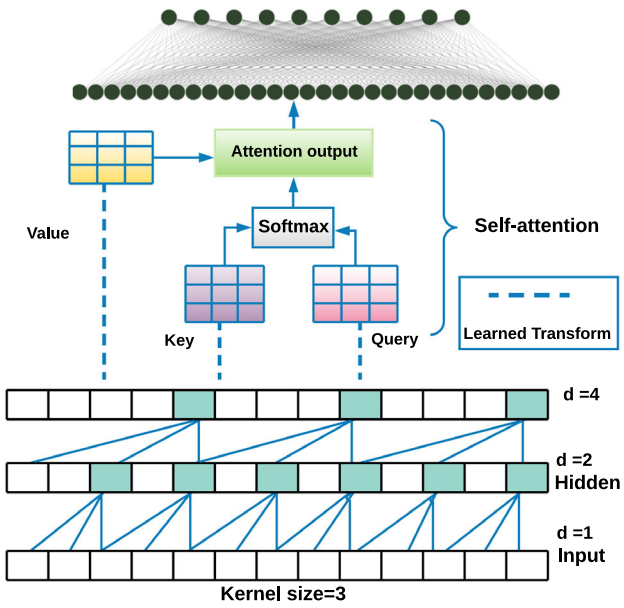

Fig. 6 Dilated causal convolution and slef-attention model

4.3 Self-attention network

The self-attention mechanism is a robust technique to compute correlation and the weighted combination between all the time steps in the input sequence [19]. After applying dilated causal convolution to render aggregated multi-scale contextual information, multi-headed self-attention is used to enable the model to focus on important and relevant time steps more than the insignificant time steps from the sequential feature maps during recognition. Hence, the attention mechanism aims to learn the most important time steps from the sequence feature maps that aid in determining more accurate recognition. Moreover, self-attention identifies relative weights for each time step in the sequence feature map by considering its similarity to all the other time-steps within the sequence. Then, the representation of each time step with relevant and important information from other time steps is transformed by the relative weights according to their importance. Self-attention mechanism has three learned linear transformation: query Q, key K, and values V, where Q and K have same vector dimension d_k, and V and outputs have same size of dimension d_v [19]. To obtain attention scores, dot product attention is applied between each query as considered to the transformed matrix of a specific time step and the key matrix of every other time step. Then the softmax function is applied on the scaled dot product value of the queries and keys to generate the attention scores. Lastly, the attention scores are used to produce a weighted representation of the value matrix for each of the time steps in the sequence. Equation (10) shows the multi-head self-attention is entirely implemented as a matrix multiplication operation.

$$f_{sa}^{(hj)}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V \tag{10}$$

The model computes the attention numerous times in parallel (multi-head) to capture distinct correlation information of the input sequence. Hence, *hj* in Eq. (10) shows output from attention head *j*, and *sa* refers to self-attention. Distinct parameters are used in Eq. (10) for computing each the key, query, and value of the n attention heads. The

outputs from the distinct multi-head attention are concatenated and transformed to the dimension of the input sequence using the learned parameter W_o as defined in Eq. (11). The outputs of the multi-head self-attention (M_{ha}) are fed into fully-connected layers, i.e., a dense layer with ReLU activation function and a soft-max layer.

$$M_{ha} = W_o \cdot \text{concat}(f_{sa}^{(h1)}, \ldots, f_{sa}^{(hn-1)}, f_{sa}^{(hn)}) \quad (11)$$

The proposed method based on dilated causal convolution foregoes recurrent architectures to accelerate the training and inference time. Causal convolution maintains the ordering of data which is crucial for HAR systems. Dilated convolution increases the receptive field and produces feature maps with multi-scale receptive fields using the different dilated rates in the convolution layers. Dilated convolution preserves the resolution of the data since the layers are dilated instead of pooling. The multi-head self-attention mechanism is employed in the proposed method to capture informative timesteps in the feature map to improve the recognition. Dilated causal convolution with a self-attention mechanism is used to make the proposed method computationally efficient and improve the result scores. Algorithm 1 besides Fig. 6 provides more information about how the layers of the proposed method are stacked.

Algorithm 1 Dilated Causal Convolution (DCC) with Multi-head Self-attention

1: **Input:** $Sensor\_data$ Sensor data
2: Generate input data using FTWs
3: First DCC layer where dilation factor = 1
4: Layer normalization and Dropout layer
5: Second DCC layer where dilation factor = 2
6: Layer normalization and Dropout layer
7: Third DCC layer where dilation factor = 4
8: Layer normalization and Dropout layer
9: Apply Multi-head Self-attention
10: Fully Connected layer
11: **Output:** Soft-Max Layer

5 Experimental setup and evaluation

In the section, we will show the details of the experimental setup and evaluation with the details of five used datasets, evaluation methods and results.

5.1 Datasets and preprocessing

5.1.1 Ordonez smart home datasets

Human activity datasets collected in five smart homes using embedded binary sensors are used in this study to evaluate the proposed method. Ordóñez home A and B [50] are two real-world smart homes that can record human daily physical activities using non-intrusive binary sensors. Different binary sensors are used in these two smart homes to detect different human activities. For example, passive infrared (PIR) sensors are used to detect human movements in a limited area. Pressure sensors on beds and couches are used to detect the user's presence. Reed switches on cupboards and doors are used to measure open or close status, and float sensors in the bathroom to measure toilet being flushed or not. Table 1 shows details about the residents, sensors, and the number of activities of the Ordóñez smart homes A and B. In Ordóñez smart home A, twelve binary sensors were used to record nine human activities in fourteen days over a period of 20,358 min. In Ordóñez smart home B, twelve binary sensors were used to record ten human activities in twenty-two days over a period of 30,469 min. The common activities from Ordóñez homes A and B are *Breakfast, Lunch, Sleeping, Grooming, Leaving, Idle, Snack, Showering, Spare Time/TV, and Toileting*, respectively. In addition to these activities, Ordóñez home B has the activity *Dinner*.

5.1.2 Kasteren smart home datasets

Kasteren home A, B, and C datasets were recorded from other three different smart homes using non-intrusive and embedded binary sensors as well [51]. Table 1 also shows the details of these three datasets regarding to the residents, the number of sensors and activities. In Kasteren home A, fourteen sensors used to record ten human activities in 25 days over a period of 40,005 min. In Kasteren home B, twenty three binary sensors used to record thirteen human activities in 14 days over a period of 38,900 min. In Kasteren home C, twenty one binary sensors used to record sixteen human activities in nineteen days over a period of 25,486 min.

5.1.3 Wearable smartphone (inertial sensors) dataset

Dataset for human activity recognition was build by recording activities of daily living (ADL) of 30 study participants while carrying a waist-mounted smartphone with embedded inertial sensors [52, 53]. The participants within an age bracket of 19–48 years performed six daily activities in which three activities are static postures (standing, sitting, lying), and three activities are dynamic activities (walking, walking downstairs, and walking upstairs). The participants wore a smartphone (Samsung Galaxy S II) on the waist to record the activities. Embedded accelerometer and gyroscope were used to capture 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50 Hz. The activities were video-recorded

Table 1 Details of the 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 (days)	14	21	25	14	19

Table 2 Frequency of activities in the Ordonez datasets

Activity	Home A	Home B
Leaving	1664	5268
Breakfast	120	309
Toileting	138	167
Spare Time/ TV	8555	8984
Dinner	–	120
Sleeping	7866	10,763
Snack	6	408
Grooming	98	427
Showering	96	75
Idle	1598	3553
Lunch	315	395
Total	20,456	30,427

Table 4 Frequency distribution of activities in the Wearable smartphone (inertial sensors) dataset

Activity	Training samples	Testing samples
Walking	1226	496
Walking_upstairs	1073	471
Walking_downstairs	986	420
Sitting	1286	491
Standing	1374	532
Laying	1407	537

to manually annotate the dataset. The dataset is randomly split into a training set with 70% of participants' data and a testing set with 30% of participants' data. Participants performed six activities: (i) Walking; (ii)

Table 3 Frequency of activities in the Kasteren datasets

Activity	Home C	Activity	Home B	Activity	Home A
Eating	345	Brush_teeth	25	Idle	7888
Idle	5883	Eat_brunch	132	Brush_teeth	21
Brush_teeth	75	Eat_dinner	46	Get_drink	21
Get_dressed	70	Get_a_drink	6	Get_snack	24
Get_drink	20	Get_dressed	27	Go_to_bed	11599
Get_snack	8	Go_to_bed	6050	Leave_house	19693
Go_to_bed	7395	Idle	20,049	Prepare_Breakfast	59
Leave_house	11,915	Leaving_the_house	12,223	Prepare_Dinner	325
Prepare_Breakfast	78	Prepare_brunch	82	Take_shower	221
prepare_Dinner	300	Prepare_dinner	87	Use_toilet	154
Prepare_Lunch	58	Take_shower	109	–	–
Shave	57	Use_toilet	39		
Take_medication	6	Wash_dishes	25	–	–
Take_shower	184		–	–	
Use_toilet_downstairs	57	–	–	–	–
Use_toilet_upstairs	35	–	–	–	–
Total	26,486	Total	38,900	Total	40,005

Table 5 Frequency distribution of wearable wireless identification and sensing datasets

Activity	RoomSet1	RoomSet2
Sit on bed	15,162	1244
Sit on chair	4381	530
Lying	30,983	20,537
Ambulating	1956	335

Walking_upstairs; (iii) Walking_downstairs; (iv) Sitting; (v) Standing; (vi) Laying. Table 4 shows the frequency distribution of activities in the training and testing sets. The accelerometer and gyroscope signals were preprocessed using noise filters. Furthermore, the signals were sampled in fixed-width sliding windows of 2.56 s and 50% overlap.

5.1.4 Wearable wireless identification and sensing data

Fourteen elderly volunteers from 78 to 78 ± 4.9 years old wore Wearable Wireless Identification and Sensing Platform (W^2ISP) tag [54–56]. The W^2ISP placed on top of their garment at the sternum level to capture trunk movements and recognize activities: (i) sit on bed; (ii) sit on chair; (iii) lying; (iv) ambulating. The activities were performed in two clinical room configurations (*Roomset1* and *Roomset2*) for ambulatory monitoring of older patients. Table 5 shows the frequency distribution of activities from both datasets:Roomset1 and Roomset2.

5.1.5 Preprocessing smart home data

The timeline of the human daily activities for all the smart homes data is segmented in time slots using the window size Δt = 1 min. The raw sensor data from smart homes provide the start time and end time of the sensor activations as well as the type (such as pressure sensor), location (such as bed), and place (such as bedroom) of the sensors. To generate the input datasets by preprocessing the raw sensor data, multiple and incremental fuzzy temporal windows (FTW) are used. FTW is used as a successful technique to segment the sensor data and prepare the input datasets [4, 6, 9, 18, 57]. FTW has shown that it can capture signal sensors of a long and short duration of human activities such as sleep or snack from raw sensor data [4, 57]. This increases the recognition results of the temporal models. Furthermore, temporal models i.e., LSTM and 1D CNN achieved better recognition results for activity recognition when the input datasets are generated by FTW compared to other methods such as Equally Sized Temporal Windows (ESTWs), Raw and Last Activation (RLA), and Raw and Last Next Activation (RLNA) [4, 6].

5.2 Models hyper-parameters

In this section, the parameters of all the models in this study are shown. A range of the following parameters used in a series of trial and error experiments over these ranges to find optimal parameters.

- Learning rates from 0.0001 to 0.01.
- Batch sizes values 32, 64, 128, and 256
- Dropout rate values 20%, 30%, 40%, and 50%.
- Number of epochs from 1 to 100.

Based on the series of trial and error experiments, we observed that 0.001 for the learning rate, 64 for the batch size with a 20% dropout rate with 50 epochs are the most appropriate hyper-parameters for the models to converge. To find a proper number of epochs, early stopping as a regularization technique is used to terminate the training when validation error starts increasing. Hence, the training was stopped at the minimum of the validation loss. To find a proper learning rate over the ranges in experiments, other hyper-parameters were fixed. This process is repeated until all the hyper-parameters are set. A large batch size can make training faster and require more memory space [6]. On the contrary, smaller batch size requires less memory space with slightly slower training but can cause the model to converge quickly, hence it is mostly a trade-off problem [6]. The 20% dropout rate is used to prevent the models from overfitting as a regularization technique [58]. The dropout technique ignores randomly selected neurons during the training process. The dropout technique temporally disconnects the ignored neurons on the forward pass hence in the backward pass their weights will not be updated. Layer normalization that normalizes the input data across the features is used after each dilation causal convolution [59]. Layer normalization can reduce the training time as empirically shown in [59].

5.3 Measure evaluation

F1-score as a metric is used to compare the performance of the proposed approach with other temporal methods. Accuracy is often used to evaluate the performance of classifiers. However, accuracy in the presence of imbalanced classes cannot be an appropriate measure for classification because less presented classes have a very little impact on accuracy as compared to the prevalent classes [6]. Hence, F1-score is employed to measure and evaluate all the temporal models since F1-score is the weighted average of recall and precision that can provide more insight into the functionality of the temporal models than the accuracy metric [4]. F1-score is calculated in Eqs. (12) and (13).

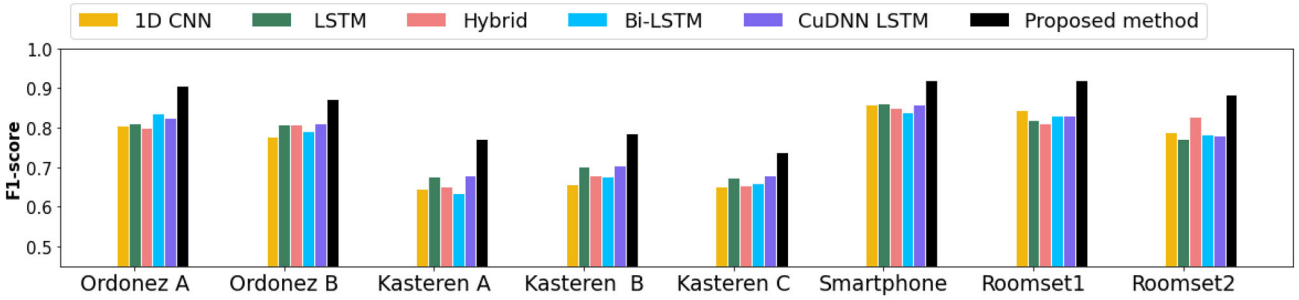

Fig. 7 Average F1-score of proposed method compared with the state-of-the-art techniques from eight the datasets

Table 6 Results of F1-score and training time in seconds from Ordonez Home A dataset

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed method
Breakfast	82.74	80.19	84.65	85.65	79.98	83.11	85.71
Grooming	46.66	57.14	51.28	74.21	62.19	75.32	80.01
Leaving	97.20	97.28	96.43	96.11	96.77	95.29	99.75
Lunch	95.65	96.92	94.87	95.42	95.34	95.44	96.93
Showering	78.94	75.45	77.94	79.42	78.12	80.65	93.84
Sleeping	96.77	96.34	96.63	95.57	94.89	97.53	97.63
Snack	64.66	67.22	55.83	67.02	69.99	70.74	84.82
Spare time	98.50	98.04	97.84	96.66	98.81	96.83	98.57
Toileting	63.75	61.09	64.71	62.71	67.42	69.89	79.76
Average	80.54	81.07	80.02	83.64	82.61	84.97	90.78
Training time	96.81	984.54	631.51	1890.62	250.814	1012.42	148.26

Table 7 Results of F1-score and training time in seconds from Ordonez Home B dataset

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed method
Breakfast	75.12	64.44	69.38	68.93	66.83	74.87	76.87
Grooming	65.75	85.55	85.36	86.12	81.62	85.33	88.87
Leaving	88.34	93.05	90.87	89.52	92.86	89.79	93.33
Lunch	98.95	81.18	77.00	79.68	83.21	95.21	99.63
Showering	78.94	79.91	78.56	77.69	80.78	79.43	82.84
Sleeping	98.11	85.71	86.76	83.29	86.82	96.37	98.30
Snack	67.92	75.86	73.41	75.42	73.21	76.16	78.59
Spare time	74.63	78.51	77.24	73.32	77.98	78.21	81.48
Toileting	48.91	80.00	83.62	76.47	83.32	83.56	86.11
Dinner	82.23	84.78	85.32	82.51	85.49	86.19	89.45
Average	77.89	80.89	80.75	79.29	81.21	84.51	87.34
Training time	201.76	1451.22	852.16	2548.01	379.057	1241.42	271.89

$$\textbf{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{12}$$

$$\textbf{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \textbf{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

where TP, FP, FN are the number of true positives, false positives, and false negatives, respectively. Moreover, F1-score is widely used in activity recognition [4, 6, 18, 35].

Table 8 F1-score results and training time in seconds of Kasteren smart home A dataset

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed method
Brush_teeth	20.22	24.08	37.86	21.56	31.46	43.59	54.44
Get_drink	51.76	56.84	48.87	42.81	57.21	59.33	66.92
Get_Snack	50.00	53.21	51.23	55.71	56.42	57.22	63.69
Go_to_bed	79.72	74.63	73.21	78.8	73.31	80.16	86.54
Leave_house	79.80	81.58	80.28	76.37	78.89	80.02	84.28
Prepare_breakfast	76.66	74.51	72.41	74.95	75.57	76.97	83.32
Prepare_Dinner	83.20	85.24	80.39	87.94	86.48	89.56	95.42
Take_shower	84.37	81.43	79.71	74.86	83.69	85.13	89.11
Use_toilet	56.60	66.11	63.06	58.42	67.85	67.82	71.97
Average	64.70	67.59	65.22	63.49	67.98	71.09	77.29
Training time	106.89	1381.49	975.21	2137.95	497.916	1056.29	148.05

Table 9 Results of F1-score and training time in seconds of Kasteren smart home B datasets

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed methods
Brush_teeth	23.10	37.62	33.25	32.57	39.55	42.89	51.18
Eat_brunch	88.42	90.14	87.53	91.72	89.87	90.93	95.92
Eat_dinner	83.19	85.23	86.68	86.01	86.31	86.79	90.02
Get_a_drink	17.84	31.18	22.34	25.61	33.03	44.15	53.00
Go_to_bed	95.11	99.01	99.21	98.91	97.94	96.32	99.73
Leaving_the_house	91.13	91.75	86.14	87.46	92.00	92.98	96.39
Prepare_brunch	77.48	80.19	83.11	85.92	79.96	85.62	88.10
Get_dressed	16.66	22.58	20.08	27.10	23.41	31.79	42.63
Prepare_dinner	93.11	97.29	94.90	97.00	96.87	96.21	97.51
Take_shower	76.82	79.12	82.71	75.91	78.91	81.95	83.13
Use_toilet	47.78	52.51	47.08	55.71	53.22	56.13	62.18
Wash_dishes	76.61	76.12	73.19	49.28	75.80	75.38	82.36
Average	65.63	70.22	68.01	67.76	70.57	73.42	78.51
Training time	99.16	1189.38	781.41	1983.90	478.563	902.14	137.35

5.4 Results and discussion

In this section, the experimental results of the proposed dilated causal convolution with the self-attention model for HAR are presented and discussed. The achieved results of each activity based on multiple models compared with the proposed are presented. Besides, the training time of all the temporal models is shown to be easily compared with the training time of the proposed method. The results of the proposed method are compared with temporal models: 1D CNN, LSTM, hybrid 1D CNN + LSTM, CudNNLSTM, and Bidirectional LSTM. The proposed method improved the results of HAR by 5% up to 7% compared with LSTM, 1D CNN, hybrid 1D CNN + LSTM, CudNNLSTM, and Bidirectional LSTM and reduced the training time. Figure 7 shows the results of the proposed method compared to the state-of-the-art techniques on eight datasets. The results indicate that the proposed method outperformed the temporal and recurrent-based models for human activity recognition from all the datasets.

5.4.1 Results from Ordóñez datasets

Tables 6 and 7 show the F1-score and training time (seconds) of the proposed method against the temporal models from Ordóñez smart homes A and B datasets. The results show that the proposed method outperforms the temporal models (LSTM, 1D CNN, hybrid 1D CNN + LSTM, CudNNLSTM, and Bidirectional LSTM). The training time in seconds is shown in Tables 6 and 7 for all the employed

Table 10 F1-score results and training time in seconds of Kasteren home C datasets

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed methods
Eating	76.71	81.32	79.69	80.18	80.36	80.98	85.31
Brush_teeth	51.27	61.56	62.82	60.73	62.59	63.55	68.11
Get_dressed	53.47	55.90	51.47	54.78	56.32	56.82	61.17
Get_drink	42.13	47.61	50.40	38.99	47.91	48.11	51.71
Get_snack	64.14	67.74	65.53	68.16	68.39	67.86	72.23
Go_to_bed	94.86	95.11	91.48	94.21	96.04	95.41	96.12
Leave_house	93.81	90.18	92.74	89.05	91.52	92.57	94.17
Prepare_Breakfast	76.35	75.74	73.15	77.78	76.81	78.45	83.42
Prepare_Dinner	77.01	79.74	68.32	71.53	78.49	79.68	84.29
prepare_Lunch	74.12	76.21	77.21	73.55	77.07	78.39	85.73
Use_Toilet_Downstairs	42.68	40.90	41.46	37.98	41.69	45.17	59.29
Use_toilet_upstairs	35.21	43.27	30.97	45.57	45.32	46.21	51.19
Shave	73.83	75.32	77.15	71.73	76.42	78.25	81.01
Take_medication	48.37	43.74	45.32	49.32	42.39	45.21	57.09
Take_shower	72.18	75.29	74.34	75.87	75.71	75.42	78.16
Average	65.02	67.30	65.47	65.96	67.80	68.79	73.93
Training time	84.42	954.32	545.49	1652.71	332.765	789.35	95.14

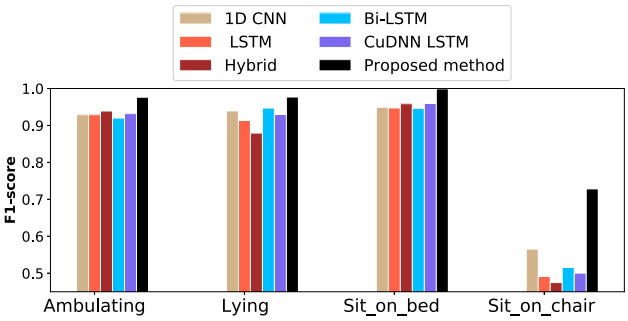

Fig. 8 F1-score from wearable dataset of RoomSet1

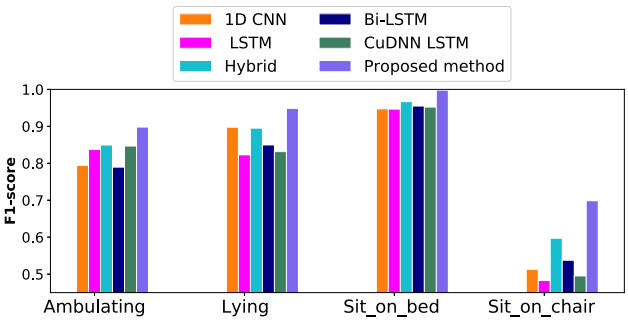

Fig. 9 F1-score from wearable dataset of RoomSet2

methods. The training time of the proposed method is much lower compared to the training time of LSTM, hybrid 1D CNN+LSTM, and Bidirectional LSTM with slightly higher training time than the 1D CNN training time. This indicates that the proposed method reduced the training time and improved the HAR systems significantly from Ordóñez smart homes. Importantly the proposed method accelerated the training time compared to the CudNNLSTM model which is a fast LSTM version and backed by Cuda library.

5.4.2 Results from Kasteren datasets

Tables 8, 9 and 10 show the results of the proposed method compared to the temporal models (LSTM, 1D CNN, hybrid 1D CNN + LSTM, CudNNLSTM, and Bidirectional LSTM) from Kasteren smart homes A, B, and C, respectively. The results of the F1-score show that the proposed method improved HAR from Kasteren datasets. The results show that the result scores are improved for each activity and the average result score. The proposed method considerably reduced the training time compared with the recurrent neural network-based architecture methods with reasonably higher training time than the 1D CNN training time. The results indicate that dilated causal convolution with self-attention can effectively improve the performance of HAR systems and reduce the training time.

5.4.3 Results from wearable sensors datasets

Figures 8, 9, and 10 show the results of the experiments that achieved based on wearable sensors for HAR. The results of the proposed method compared to the results of the temporal models (LSTM, 1D CNN, hybrid 1D

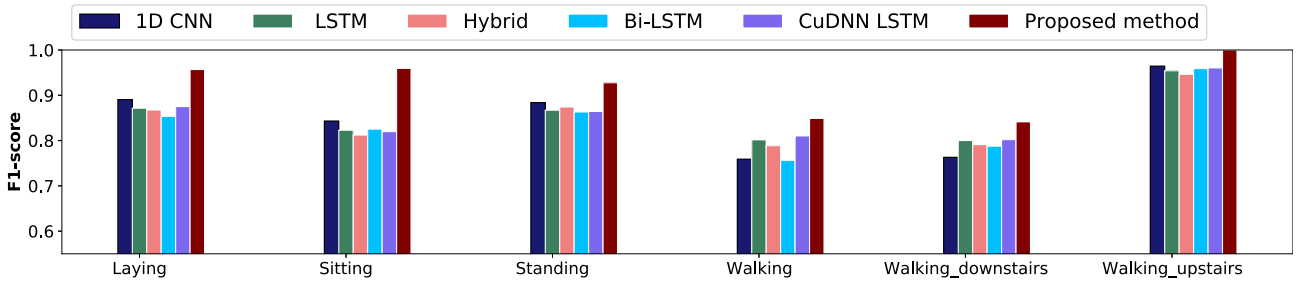

Fig. 10 F1-score from wearable smartphone dataset

Table 11 Results of F1-score and training time in seconds from smartphone dataset

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed method
Laying	89.03	87.14	86.76	85.35	87.51	89.67	95.69
Sitting	84.32	82.29	81.22	82.53	81.98	86.45	95.94
Standing	88.38	86.71	87.42	86.32	86.43	88.92	92.77
Walking	75.90	80.17	78.87	75.64	81.03	80.89	84.89
Walking_downstairs	76.31	80.01	79.11	78.76	80.21	80.11	84.14
Walking_upstairs	96.44	95.42	94.63	95.89	96.05	96.93	100.00
Average	85.89	86.14	85.00	84.08	86.03	87.16	92.24
Training time	92.45	547.62	463.64	874.35	212.63	697.89	121.13

Table 12 Results of F1-score and training time in seconds from wearable dataset of RoomSet1

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed method
Ambulating	92.91	92.58	93.67	92.02	93.19	93.67	97.63
Lying	93.94	91.34	87.94	94.71	92.97	94.12	97.70
Sit_on_bed	94.91	94.74	95.89	94.64	95.94	95.31	99.90
Sit_on_chair	56.51	49.12	47.48	51.58	50.04	60.52	72.84
Average	84.56	81.94	81.24	83.23	83.03	85.90	92.02
Training time	89.11	406.95	389.21	944.42	231.56	618.64	198.24

Table 13 Results of F1-score and training time in seconds from wearable dataset of RoomSet2

Activity	CNN	LSTM	Hybrid	Bi-LSTM	CuDNN LSTM	DeepConvLSTM + Attention	Proposed method
Ambulating	79.42	83.75	84.95	78.95	84.67	85.43	89.79
Lying	89.75	82.29	89.50	84.97	83.16	89.42	94.85
Sit_on_bed	94.74	94.66	96.75	95.49	95.21	96.21	99.79
Sit_on_chair	51.31	48.27	59.70	53.75	49.51	62.56	69.87
Average	78.80	77.24	82.72	78.24	78.13	83.40	88.57
Training time	36.30	91.64	85.94	158.11	78.42	135.31	56.05

CNN + LSTM, CudNNLSTM, and Bidirectional LSTM). Tables 11, 12, and 13 show the detailed results and the training time of all the models. Table 11 particularly shows the results of the experiments obtained based on smartphone sensors data. The results of the wearable sensors data demonstrate the outstanding performance of our

Table 14 Results of F1-score of ablation studies of the proposed method

Datasets	Without dilated convolution	Without attention	Without Causal	Without all	Proposed method
Ordonez Home A	84.93	83.24	85.41	80.54	90.78
Ordonez Home B	83.79	81.11	84.19	77.89	87.34
Kastern Home A	72.87	69.78	74.21	64.70	77.29
Kastern Home B	74.68	68.81	75.11	65.63	78.51
Kastern Home C	70.44	67.13	70.88	65.02	73.93
Smartphone dataset	87.13	86.27	89.25	85.89	92.24
Wearable RoomSet1	87.17	84.29	88.72	84.56	92.02
Wearable RoomSet2	84.09	82.25	85.19	78.80	88.57

proposed method compared to the state-of-the-art techniques. The training time of the proposed method is shorter than the training time of all the temporal and recurrent models except the training time of 1D CNN. The proposed method improved the performance of each activity as well as the average performance of all activities compared to recurrent and temporal models based on all wearable sensor data.

5.4.4 Proposed method compared to the DeepConvLSTM + attention

Results of our proposed method are compared with the results achieved by the DeepConvLSTM + Attention [13] for all the datasets. Results of the DeepConvLSTM + Attention are shown from all the datasets in this research and the training time. Since the DeepConvLSTM + Attention works based on the combination of 2D CNN and LSTM with an attention mechanism, it requires more time to process the input data compared to our proposed method. Moreover, compared to DeepConvLSTM + Attention, our proposed method achieved better result scores with much faster training times in all the datasets. For instance, our proposed method achieved the F1-score of 90.78 and 87.51 from Ordóñez smart homes A and B datasets, respectively, while the DeepConvLSTM + Attention achieved the F1-score of 84.97 and 84.51 for the same datasets with higher training times.

Our proposed method dispenses the recurrence setting entirely to accelerate the training time and boost the performance of HAR systems. Dilated convolution aggregates multi-scale contextual information to render informative feature maps. Causal convolution in the proposed method ensures the model cannot violate the ordering of the sequential temporal input data. The proposed method can focus on the important timesteps using the attention mechanism to improve the recognition process. The proposed method improved the results of each activity in addition to the average results of all the activities and all the datasets.

5.4.5 Ablation study of the proposed method

Ablation studied is conducted to show performance of the proposed method without dilated convolution, causal convolution and attention mechanism. Table 14 shows the results of these models and the results of the proposed method without these three techniques as well as the results of the proposed method from all the datasets. The results show that how the proposed method is affected by each of the dilated convolution, causal convolution and attention mechanism. For example, the proposed method achieved the F1-score of 90.78, while the proposed method without dilated convolution achieved the F1-score of 84.93, without attention achieved the F1-score of 83.24, without causal convolution achieved the F1-score of 85.41. Moreover the proposed method without these three techniques achieved the F1-score of 80.54. The proposed method without using attention mechanism has achieved lowest results scores from all the datasets compared to the proposed method without using dilated and causal convolutions. Hence, the results indicated that the attention mechanism has a higher contribution in the proposed method compared to dilated and causal convolutions. Beside the ablation study, the proposed method is compared with the DeepConvLSTM + Attention method and many temporal and recurrent models: LSTM, 1D CNN, hybrid 1D CNN + LSTM, CudNNLSTM, and Bidirectional LSTM.

6 Conclusion

This study proposes dilated causal convolution with multi-head self-attention to accelerate training time and improve the performance of HAR systems from smart home and wearable sensor data. Thorough experiments are conducted on eight real-world smart home and wearable datasets to evaluate the proposed method against the temporal and recurrent-based architecture methods. The results of the experiments show that the proposed method significantly improved the accuracy of HAR and reduced the training time compared to the state-of-the-art techniques. The proposed method improved the performance of HAR systems by up to 7% compared with LSTM, 1D CNN, hybrid 1D CNN + LSTM, CuDNNLSTM, and Bidirectional LSTM using wearable sensors and smart home sensors data.

The operation of the self-attention mechanism scales quadratically with the input sequence length which can increase training time because it appends more weight parameters to the model. To address this limitation, our future work will investigate a newly proposed method in human activity recognition to further accelerate the training time and enhance the performance of HAR by introducing a lightweight multi-head self attention mechanism.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. 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 (2018)
2. 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, pp 719–722. IEEE
3. Lee D, Helal S (2013) From activity recognition to situation recognition. In: International conference on smart homes and health telematics, pp 245–251. Springer
4. Javier M-Q, Shuai Z, Chris N, 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
5. 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), pp 1839–1845. IEEE
6. 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 Inform
7. Wang W, Liu AX, Shahzad M, Ling K, Lu S (2015) Understanding and modeling of wifi signal based human activity recognition. In: Proceedings of the 21st annual international conference on mobile computing and networking, pp 65–76. ACM
8. Jindong W, Yiqiang C, Shuji H, Xiaohui P, Lisha H (2019) Deep learning for sensor-based activity recognition: a survey. Pattern Recogn Lett 119:3–11
9. Ali HR, Masashi K, Jens L (2020) Efficacy of imbalanced data handling methods on deep learning for smart homes environments. SN Comput Sci 1(4):1–10
10. Iram F, Muhammad F, Young-Koo L, Sungyoung L (2013) Analysis and effects of smart home dataset characteristics for daily life activity recognition. J Supercomput 66(2):760–780
11. Liang C, Yufeng W, Bo Z, Qun J, Vasilakos Athanasios V (2018) Gchar: an efficient group-based context-aware human activity recognition on smartphone. J Parallel Distrib Comput 118:67–80
12. Nweke HF, Teh YW, Al-Garadi MAA (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl
13. Singh SP, Lay-Ekuakille A, Gangwar D, Sharma MK, Gupta S (2020) Deep CONVLSTM with self-attention for human activity decoding using wearables. arXiv preprint arXiv:2005.00698
14. Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th annual international conference on machine learning, pp 609–616
15. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
16. Lee H, Pham P, Largman Y, Ng AY (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in neural information processing systems, pp 1096–1104
17. Zhao R, Wang J, Yan R, Mao K (2016) Machine health monitoring with LSTM networks. In: 2016 10th international conference on sensing technology (ICST), pp 1–6. IEEE
18. Ali HR, Longzhi Y, Lok WW, Wei B (2020) Joint learning of temporal models to handle imbalanced data for human activity recognition. Appl Sci 10(15):5293
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
20. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271

21. 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, pp 194–205. Springer
22. Lee S-M, Yoon SM, Cho H (2017) Human activity recognition from accelerometer data using convolutional neural network. In: 2017 IEEE international conference on big data and smart computing (bigcomp), pp 131–134. IEEE
23. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Alex G, Nal K, Andrew S, Koray K (2016) Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499
24. Pu J, Zhou W, Li H (2018) Dilated convolutional network with iterative optimization for continuous sign language recognition. In: IJCAI, vol 3, p 7
25. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122
26. Lin Zhouhan, Feng Minwei, Nogueira dos Santos Cicero, Yu Mo, Xiang Bing, Zhou Bowen, Bengio Yoshua (2017) A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130
27. Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y (2020) Deep learning for sensor-based human activity recognition: overview, challenges and opportunities. arXiv preprint arXiv:2001.07416
28. Kun X, Jianguang H, Hanyu W (2020) LSTM-CNN architecture for human activity recognition. IEEE Access 8:56855–56866
29. Bengio Y (2013) Deep learning of representations: Looking forward. In: International conference on statistical language and speech processing, pp 1–37. Springer
30. 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, pp 341–348. Springer
31. Sepp H, Jürgen S (1997) Long short-term memory. Neural Comput 9(8):1735–1780
32. Masaya I, Sozo I, Takeshi N (2018) Deep recurrent neural network for mobile human activity recognition with high throughput. Artif Life Robot 23(2):173–185
33. Hernández F, Suárez LF, Villamizar J, Altuve M (2019) Human activity recognition on smartphones using a bidirectional LSTM network. In: 2019 XXII symposium on image, signal processing and artificial vision (STSIVA), pp 1–5. IEEE
34. Ullah M, Ullah H, Khan SD, Cheikh FA (2019) Stacked LSTM network for human activity recognition using smartphone data. In: 2019 8th European workshop on visual information processing (EUVIP), pp 175–180. IEEE
35. Guan Yu, Thomas P (2017) Ensembles of deep LSTM learners for activity recognition using wearables. Proc ACM Interact Mobile Wear Ubiquit Technol 1(2):1–28
36. Zeng Y, Xiao Z, Hung K-W, Lui S (2021) Real-time video super resolution network using recurrent multi-branch dilated convolutions. Signal Process Image Commun 93:116167
37. Yingjie L (2020) Wu J (2020) A novel multichannel dilated convolution neural network for human activity recognition. Math Probl Eng
38. Chang S-Y, Li B, Simko G, Sainath TN, Tripathi A, van den Oord A, Vinyals O (2018) Temporal modeling using dilated convolution and gating for voice-activity-detection. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5549–5553. IEEE
39. Woon-Haeng H, Hyemi K, Oh-Wook K (2021) Integrating dilated convolution into dense LSTM for audio source separation. Appl Sci 11(2):789
40. Jun H, Qian Z, Liqun W, Ling P (2018) Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. IEEE Sens J 19(6):2287–2297
41. Mahmud S, Tonmoy M, Bhaumik KK, Rahman AKM, Amin MA, Shoyaib M, Asif Hossain KM, Ali AA (2020) Human activity recognition from wearable sensor data using self-attention. arXiv preprint arXiv:2003.09018
42. 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), pp 1194–1199. IEEE
43. 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
44. 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 preprint arXiv:2006.14435
45. Hammerla NY, Halloran S, Ploetz T (2016) Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint arXiv:1604.08880
46. Appleyard J, Kocisky T, Blunsom P (2016) Optimizing performance of recurrent neural networks on GPUS. arXiv preprint arXiv:1604.01946
47. Francisco Javier Ordóñez and Daniel Roggen (2016) Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1):115
48. Mike S, Paliwal Kuldip K (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
49. Alex G, Jürgen S (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5–6):602–610
50. Fco O, Paula DT, Araceli S et al (2013) Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. Sensors 13(5):5460–5477
51. van Kasteren TLM, Englebienne G, Kröse BJA (2011) Human activity recognition from wireless sensor network data: benchmark and software. In: Activity recognition in pervasive intelligent environments, pp 165–186. Springer
52. 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
53. Jorge-L R-O, Luca O, Albert S, Xavier P, Davide A (2016) Transition-aware human activity recognition using smartphones. Neurocomputing 171:754–767
54. Luis STR, Ranasinghe DC, Shi Q (2013) 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, pp 384–395. Springer
55. Shinmoto TRL, Ranasinghe DC, Shi Q, Sample AP (2013) Sensor enabled wearable RFID technology for mitigating the risk of falls near beds. In: 2013 IEEE international conference on RFID (RFID), pp 191–198. IEEE
56. Wickramasinghe A, Ranasinghe DC (2016) Recognising activities in real time using body worn passive sensors with sparse data streams: To interpolate or not to interpolate? In: Proceedings of the 12th EAI international conference on mobile and ubiquitous systems: computing, networking and services on 12th EAI international conference on mobile and ubiquitous systems: computing, networking and services, pp 21–30
57. 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
58. Nitish S, Geoffrey H, Alex K, Ilya S, Ruslan S (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

59. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv:1607.06450