

Recognizing human activities with attention mechanism on multi-source sensor data

Haixia Bi^{*ab}, Miquel Perello-Nieto^b, Emma Tonkin^b, Raul Santos-Rodriguez^b, Peter Flach^b, Ian Craddock^b

^aSchool of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China; ^bFaculty of Engineering, University of Bristol, Bristol, UK

ABSTRACT

A smart home equipped with a diversity of multimodal sensors is a meaningful setting for acquiring the health status of its residents and improving their well-being. In recent years, sensor-based activity recognition has received growing research attention. However, the multi-modal nature of these sensor platforms raises great challenges with respect to the data fusion of the different sensor sources. To solve this problem, we present an activity recognition approach incorporating attention mechanism in this paper. A Convolutional Neural Network-based training framework is developed to extract representative features for activities. Specifically, we design two attention modules-channel-wise and temporal-wise modules to capture the interdependencies between channel and temporal dimensions of its convolutional features. We evaluate the attention-based approach on a real activity recognition challenge dataset. Experiments justify that the attention network-based feature fusion can effectively improve the activity recognition performance.

Keywords: Human activity recognition, convolutional neural network, attention, fusion

1. INTRODUCTION

Currently, ageing population is becoming a severe issue for many countries in the world, bringing great pressure to the medical resources¹. Smart Home (SH) technology is a desirable approach to extend the role of healthcare management from traditional medical agencies to private homes². A smart home normally has systems with network and sensors, usually using Internet of Things (IoT) technologies³. The SH system often includes a diversity of nonobtrusive sensors, such as wearable sensors, video cameras or environmental sensors. It is therefore necessary to obtain useful health information from the sensor data collected. Under this circumstance, Human Activity Recognition (HAR) becomes a very important step to infer the behavior of home residents, providing meaningful information to the medical workers and careers^{4,5}.

HAR has attracted widespread research interests recently, and a great number of approaches have been presented to solve this issue⁶⁻¹⁰. However, the majority of these methods are proposed to handle data collected with single modality sensors. In a real smart home setting, fused sensor-based activity recognition is confronted with several challenges. In this paper, we make use of the SPHERE smart home platform⁵ as an exemplar. 1) There is high heterogeneity and uncertainty for sensors with different modalities. Almost all sensors, including wearable, environmental and video sensors, are reliant on network availability for data collection. 2) From the data fusion point of view, different sensors or features have different impacts on the activity recognition performance. Even inside the temporal window of one feature, different positions of the data sequence may have diverse influences on the outcomes of activity recognition as well. However, existing methods usually treat the features or temporal positions equally without considering the diversity. This inspires us to investigate the importance of the attention feature maps from the channel and temporal level.

Motivated by the above analysis, and inspired by Reference¹¹, we propose an attention network-based human activity recognition approach in this work. An end-to-end Convolutional Neural Network (CNN)-based training framework is developed to extract representative features for activities with fused sensor data. Specifically, an attention mechanism is introduced to put more emphasis on informative features in channel and temporal dimensions. In order to attain this goal, we apply a 1-dimensional Convolutional Block Attention Module (CBAM) with two sub-attention modules, i.e.,

* haixia.bi@xjtu.edu.cn

channel-wise and temporal-wise attention modules, to capture the interdependencies between channel and temporal dimensions of its convolutional features. We summarize the main contributions of this paper as below:

- 1) We put forward an attention-based deep learning pipeline to recognize human activities on fused sensor data. The CNN and attention mechanism are integrated as a CNN+CBAM framework to extract representative features for activities from fused sensor data.
- 2) We design multi-level and multi-dimension attention modules for time-series data, to improve representation of interests. CBAM attention modules which consists of channel-wise and temporal-wise attention modules are combined with convolution layers, by which the representation power of important features is enhanced and the unnecessary ones are attenuated.
- 3) We evaluate our framework on a real healthcare dataset. Experimental results demonstrate that the proposed framework could effectively recognize human activities based on fused sensor data, and the attention modules further boost the performance of standard CNN.

The paper is organized as follows. We introduce the proposed method in Section 2. Section 3 reports the comparative study. Section 4 summarizes the paper, and discusses the future work.

2. METHODOLOGY

2.1 Overview

We formulate the sensor-based human activity recognition as a time-series sequence classification problem in this work, where the samples are temporal sequences and the classes correspond to activities. Figure 1 illustrates the pipeline of our method. Given the multi-sensor HAR dataset, we first perform pre-processing on the original data to get a series of segments which are suitable for the deep network training. The output of the pre-processing acts as input for the consequent network, which is denoted as $\tilde{F} \in \{\tilde{f}_i \in \mathbb{R}^{T \times D}\}_{i=1}^N$, where N indicates the number of data segments, T is the time step length and D denotes the feature dimension. The full activity label set is given as $y_i \in \{1, 2, \dots, K\}$, where K is the total number of activity classes. The proposed method is designed to assign activity class label for each sample $i \in \{1, 2, \dots, N\}$.

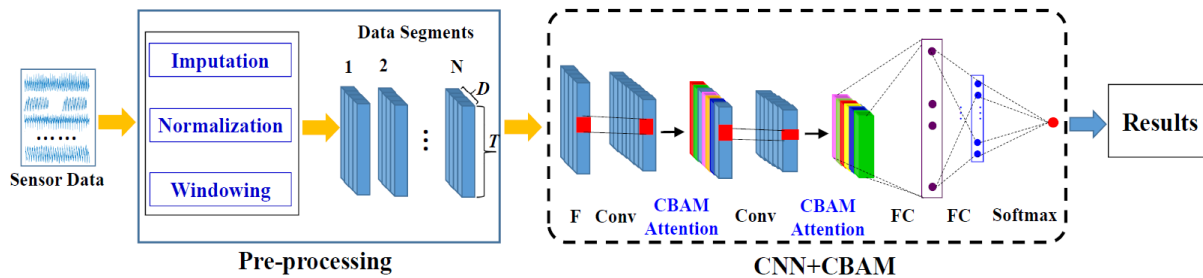


Figure 1. Method illustration.

2.2 Data pre-processing

In this paper, the SPHERE challenge dataset¹² is used as the input raw data. It was collected using a variety of heterogeneous sensors in a smart home by the SPHERE project¹³. The sensing modalities of the dataset include: wrist-worn accelerometer with received signal strength indicator (RSSI) sensors, passive environmental sensors and RGB-D cameras. The ground truth label includes 20 activities, which can be referred to in Reference¹³. There are two features which make the employed dataset challenging. 1) The sensor modalities of this dataset are complex and heterogeneous. 2) Missing data exist in several features of the dataset¹⁴⁻¹⁶.

The data pre-processing contains four steps, i.e., missing value imputation, normalization, windowing and multiple annotators processing, which will be detailed as follows.

- (1) Missing value imputation: In the SPHERE challenge dataset, missing values exist on both the RSSI sensor and video

datastreams. The missing value on the RSSI sensor is due to the lack of signal from the wearable to the receivers. For this case, we impute the missing values in RSSI by linear interpolation operation. For the video sensors, missing values emerge when the network interrupts or no human is present before the camera. Under this circumstance, we impute the missing values with 0.

(2) Normalization: In the context of data fusion, different features can have dispersed value ranges. In this work, we perform feature standardization to get standard normally distributed data by removing the mean and scaling to unit variance.

(3) Windowing: The training of the temporal deep neural networks requires data segments with fixed temporal length. For the SPHERE challenge dataset, we adopt a 2-second window without overlapping between neighboring windows in our work.

(4) Multiple annotators: In the original SPHERE challenge datasets, some annotations are probabilistic due to disagreement between different annotators. In order to get more reliable annotations, we only keep the data segments with consistent annotations between multi-annotators.

After performing the above pre-processing steps on the SPHERE challenge data, we obtain an input feature map $\tilde{F} \in \{\tilde{f}_i \in \mathbb{R}^{T \times D}\}_{i=1}^N$, where the time step T is 20, feature dimension D is 61, and there are 97,040 data segments in total.

2.3 Attention-based network: CNN+CBAM

Given the extracted feature map \tilde{F} as input, we next introduce an attention-based network, which we use CNN+CBAM to denote in this work. We illustrate the architecture in Figure 1. For the first convolutional layer, the feature maps are taken as the extracted raw features, and for the second layer, they are taken as the output of the preceding CBAM layer. The first convolutional layer contains 256 kernels with size 3, and the second convolutional layer contains 128 kernels with size 3. We set the stride of both convolution layers as 2 to downsample the feature maps. We use rectified linear units (ReLUs) as nonlinear activation functions. The first fully connected layer contains 640 units, while the second fully connected layer consists of 200 units. The Softmax layer outputs the class probabilities on the 20 activity classes. We employ cross entropy as the loss function, and ADAM optimizer¹⁷ is utilized as optimization method. We set the training batch size as 64 in the experiments, and empirically set the training epoch as 30.

We next introduce the design of the CBAM layer. CBAM was proposed in Reference¹¹ for image classification task. In this work, we revise the original CBAM layer to suit for time-series data classification. Let F denotes the input feature, CBAM block learns a 1-dimensional channel-wise attention map $M_c \in \mathbb{R}^{D \times 1}$ and a 1-dimensional temporal-wise attention map $M_t \in \mathbb{R}^{1 \times T}$, as displayed in Figure 2. The processing of CBAM can be described as:

$$F' = M_c(F) \otimes F, \tag{1}$$

$$F'' = M_t(F') \otimes F', \tag{2}$$

where \otimes indicates element-wise multiplication. The input feature F sequentially multiply with attention maps M_c and M_t , finally obtaining the output F'' .

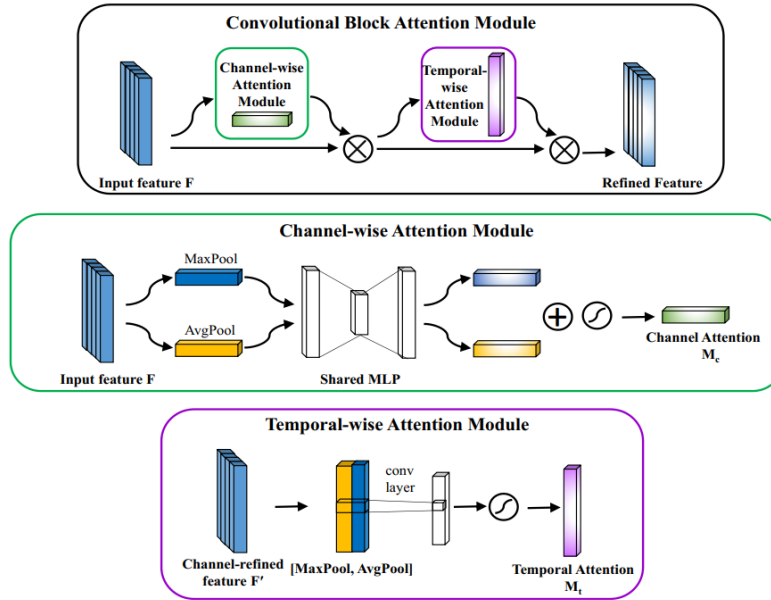


Figure 2. Diagram of CBAM and the attention sub-modules.

(1) *Channel-wise attention module.* We generate a channel-wise attention map through employing the relationship between features in different channels. The information in temporal dimension is firstly aggregated via utilizing an average pooling operation and a max-pooling operation. Afterwards, both descriptors are fed into a common network to generate the channel-wise attention map M_c . The shared network consists of an MLP (multi-layer perceptron) and a hidden layer.

(2) *Temporal-wise attention module.* The temporal-wise attention map is produced by using the relationship of features in temporal dimension. The channel information of a feature map is aggregated by means of a max-pooling operation and an average pooling operation as well, producing two 1-dimensional maps. We then concatenate the two maps and convolve them with a 1-dimensional convolution layer, which generates the 1-dimensional temporal-wise attention map finally.

3. EXPERIMENTS

In this section, we validate the effectiveness of the CNN+CBAM method by means of two groups of experiments. The first group of experiments (Experiment 1) shows the improvement on the overall classification accuracy (abbreviated as *OCA*) brought by attention modules, individually and combined.

- CNN: baseline CNN as described in Section 2.3.
- CNN+CA: baseline CNN with two channel-wise attention modules.
- CNN+TA: baseline CNN with two temporal-wise attention modules.
- CNN+2CBAMs: baseline CNN with two CBAM modules.

The second group (Experiment 2) compares the *OCA* values with respect to the number and position of different CBAM modules.

- CNN: baseline CNN.
- CNN+1stCBAM: CBAM module after the first convolution layer.
- CNN+2ndCBAM: CBAM module after the second convolution layer.

- CNN+2CBAMs: CBAM modules after both convolution layers.

The dataset is split into two subsets, a training subset with 80% of the samples and a validation subset with the remaining 20% of the samples in a stratified manner. We run 10 independent repetitions on all experiments obtaining new training and validation partitions, and then average them as the final results. All the experiments are implemented in Pytorch framework on a PC which is equipped with GeForce RTX 3090 GPU and 16-GB of RAM.

The baseline CNN architecture was chosen based on the validation performance of a group of architectures with different numbers of feature maps, and we selected the one with the best validation performance. The other compared approaches are with the same baseline architecture without further hyper-parameter tuning. The attention modules use the default parameter setting as specified in Reference¹⁰. It should be noted that the CNN+CBAM framework could be applied on different architectures of CNN. However, we only present the results on the CNN architecture as described in Section 2.3.

Figure 3 illustrates the *OCA* curves with respect to the epoch number, where Figure 3a shows the improvement brought by channel-wise and temporal-wise attention modules, while Figure 3b displays the results using different numbers and positions of CBAM modules. Tables 1 and 2 show the numerical results of the compared methods respectively.

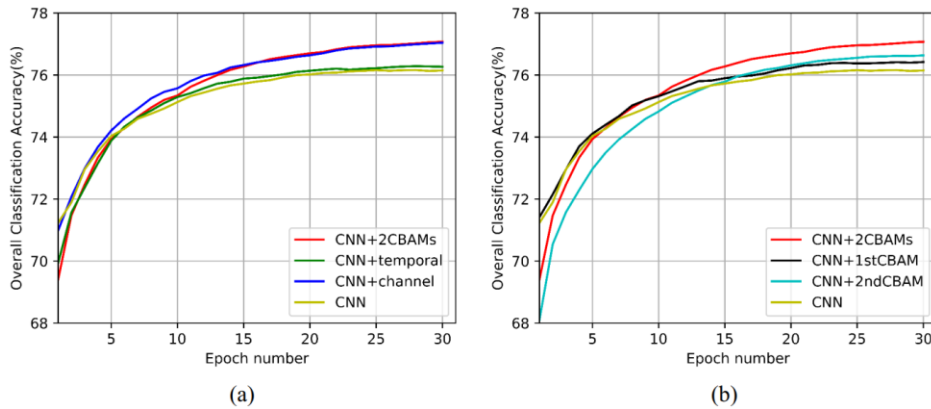


Figure 3. (a): Validation performance comparison between CNN, CNN+channel, CNN+temporal and CNN+2CBAMs (Experiment 1); (b) Validation performance comparison between different number of CBAM modules (Experiment 2).

Table 1. *OCA* values (%) for Experiment 1.

Method	OCA value
CNN	76.14
CNN+CA	77.04
CNN+TA	76.25
CNN+CBAM	77.07

Table 2. *OCA* values (%) for Experiment 2.

Method	OCA value
CNN	76.14
CNN+CA	76.41
CNN+TA	76.63
CNN+CBAM	77.07

From Figure 3a, we see that both the channel-wise attention module and temporal-wise attention module contribute to the improvement of the classification accuracy. The channel-wise attention module brings more improvement of *OCA* value than the temporal-wise attention module. This may be because the data instances we use are segmented by time sequence. However, the starting or ending of actual activities are not necessarily related to the time sequence, which means there is no clear relationship pattern between the time position and activities we study. The channel in our experiments correspond to sensor modalities. The accelerometer and RSSI signals are collected via the wristband wearable device, which are tightly related to the activities of participants. However, the PIR sensors and video sensors only have values when human are present before these sensors. In addition, the data collected by PIR sensors are noisy, because the PIR sensor sometimes reports false positive values in some cases. The channel-wise attention module learns the importance of different channels, and assigns higher weight on the important feature maps, and depress the unimportant ones, which explains the larger improvement brought by channel-wise attention module. From Table 1, we can find the channel-wise attention module enhances the *OCA* value by 0.90%, while the temporal-wise attention module improves the *OCA* value by 0.11%. The CNN+CBAM which contains both attention sub-modules outperforms the baseline method by 0.93%.

From Figure 3b and Table 2, we can find that the position and number of CBAM modules influence the *OCA* value. Incorporating two CBAM modules achieves better classification accuracy compared with one CBAM module. When only one CBAM module is integrated, the CBAM module located in the higher layer contributes more to the classification performance. This is because the higher level features are more complex and abstract than the raw features. Therefore, they are more informative for the network to decide which channel or temporal position to choose with respect to attention. From Table 2, we can observe that the first CBAM module improves the *OCA* value by 0.27%, while combining the second CBAM module promotes the *OCA* value by 0.49%.

4. CONCLUSION

This paper presented a CNN pipeline with attention mechanism for recognizing human activity recognition in sensor fusion scenario. The pre-processing step of the pipeline involves missing value imputation, feature-wise normalization and windowing, which generates usable data segments for the network training. We integrated multi-level CBAM modules with a CNN to extract discriminative features for activities. Each CBAM module contains a channel-wise attention module and a temporal-size module, which are devised to capture the interdependencies between channel and temporal dimensions of their convolutional features. We performed a series of experiments, whose results on the real SPHERE dataset justify that the proposed method could effectively recognize human activities based on fused sensor data, and the attention modules further enhance the performance of deep neural network. In the future, we will develop joint activity and localization recognition using attention-enhanced network based on fused sensor data.

REFERENCES

- [1] Chen, L., Nugent, C. D. and Wang, H., "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering* 24(6), 961-974 (2011).
- [2] Chan, M., Est'ève, D., Escriba, C. and Campo, E., "A review of smart homes—present state and future challenges," *Computer Methods and Programs in Biomedicine*, 91(1), 55-81 (2008).
- [3] Alaa, M., Zaidan, A. A., Zaidan, B. B., Talal, M. and Kiah, M. L. M., "A review of smart home applications based on internet of things," *Journal of Network and Computer Applications*, 97, 48-65 (2017).
- [4] Diethe, T., Twomey, N., Kull, M. and Flach, P., "Craddock I: Probabilistic sensor fusion for ambient assisted living," *arXiv preprint arXiv:1702.01209*, (2017).
- [5] Woznowski, P., Burrows, A., Diethe, T., Fafoutis, X., Hall, J., Hannuna, S., Camplani, M., Twomey, N., Kozłowski, M., Tan, B., et al., "SPHERE: A sensor platform for healthcare in a residential environment," *Designing, Developing, and Facilitating Smart Cities*, 315-333 (2017).
- [6] Bi, H., Perello-Nieto, M., Santos-Rodriguez, R. and Flach, P., "Humanactivity recognition based on dynamic active learning," *IEEE Journal of Biomedical and Health Informatics* 25(4), 922-934 (2021).
- [7] Bi, H., Perello-Nieto, M., Santos-Rodriguez, R., et al., "An active semi-supervised deep learning model for human activity recognition," *Journal of Ambient Intelligence and Humanized Computing*, 1-17 (2022).
- [8] Tang, Y., Zhang, L., Min, F., et al., "Multi-scale deep feature learning for human activity recognition using wearable sensors," *IEEE Transactions on Industrial Electronics*, (2022). DOI:10.1109/TIE.2022.3161812
- [9] Qiu, S., Zhao, H., Jiang, N., et al., "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion* 80, 241-265 (2022).

- [10] Tonkin, E. L., Nieto, M. P., Bi, H. and Vafeas, A., "Towards a methodology for acceptance testing and validation of monitoring bodyworn devices," IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 1-6 (2020).
- [11] Woo, S., Park, J., Lee, J. Y., et al., "CBAM: Convolutional blockattention module," Proceedings of the European Conference on Computer Vision (ECCV), 3-19 (2018).
- [12] Twomey, N., Diethel, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P., et al., "The SPHERE challenge: Activity recognition with multimodal sensor data," arXiv preprint arXiv:1603.00797, (2016).
- [13] Zhu, N., Diethel, T., Camplani, M., Tao, L., Burrows A, Twomey N, Kaleshi D, Mirmehdi M, Flach P, Craddock I: Bridging e-health and the internet of things: The sphere project. IEEE Intelligent Systems 30(4), 39–46 (2015)
- [14] Bi, H., Xu, F., Wei, Z., Xue, Y. and Xu, Z., "An active deep learning approach for minimally supervised PolSAR image classification," IEEE Transactions on Geoscience and Remote Sensing 57(11), 9378-9395 (2019).
- [15] Bi, H., Xu, L., Cao, X., Xue, Y. and Xu, Z., "Polarimetric SAR image semantic segmentation with 3d discrete wavelet transform and Markov random field," IEEE Transactions on Image Processing, 29, 6601-6614 (2020).
- [16] Bi, B., Yao, J., Wei, Z., Hong, D. and Chanussot, J., "PolSAR image classification based on robust low-rank feature extraction and Markov random field," IEEE Geoscience and Remote Sensing Letters 19, 1-5 (2020).
- [17] Kingma, D. P. and Ba, J., "ADAM: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).