Journal of Advanced Marine Engineering and Technology, Vol. 47, No. 3, pp. 131~136, 2023
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2023.47.3.131

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

A study on light weight of radar-based human activity recognition model using self-attention mechanism

Seong-Beom Jeong[1] · Hong-Il Seo[2] · Gi-Do Choi[3] · Dong-Hoan Seo[†]

(Received May 9, 2023 ; Revised May 22, 2023 ; Accepted May 30, 2023)

Abstract: Recently, radar-based human activity recognition (HAR) technology has been actively studied in the field of artificial intelligence by applying radar datasets to deep learning (DL) models to automatically learn and classification. This is one of the important applications in the field of activity recognition and can be used in various fields, such as exercise tracking, smart homes, self-driving cars, and health status monitoring, by recognizing human daily activity patterns. However, DL models are complex and have significant computational costs and numerous parameters to process and classify high-dimensional radar datasets. Therefore, their implementation on commercial mobile devices is limited by their computational complexity. Therefore, in this study, we propose a lightweight HAR DL model using Self-Attention technology to solve the complexity and computational costs of these DL models. Experimental results demonstrate that this model can maintain its performance while reducing the number of parameters required. In the future, these lightweight models will not only be usable on mobile devices but will also require lower computing power and memory capacity; therefore, they are expected to be used in various fields as time series-based DL models.

Keywords: Human activity recognition, Radar signal processing, CNN, LSTM, Self-attention

1. Introduction

Human activity recognition (HAR) technology is a research topic that has received considerable attention in recent years due to its potential applications in various fields. This technology can be used in various fields such as health status monitoring, exercise tracking, smart homes, and self-driving cars by recognizing daily human activity patterns **[1]**. The three basic types of HAR—camera-based **[2]**, wearable sensor-based **[3]**, and radar-based **[4]** systems—each have their strengths and weaknesses.

Camera-based HAR systems use optical sensors to capture images or videos to analyze human behavior. This method is convenient because it does not require the subject to wear a separate device. However, camera-based systems are sensitive to lighting conditions, and privacy concerns can arise when these systems capture visual data. Wearable sensor-based HAR systems rely on body-mounted devices such as smartphones or smartwatches to collect data on acceleration, gyroscopic motion, and geomagnetic information. These types of systems provide a high degree of accuracy and detailed insights into user activity. However, wearing or carrying these devices all the time may be inconvenient to the user.

Radar-based HAR systems utilize the micro-Doppler effect to analyze the movement, vibration, and rotation of the subject's limbs. This system can provide information regarding the range and speed of an object without requiring a wearable device. In addition, radar-based systems are passive and noninvasive, giving them a better privacy advantage than camera-based systems. However, the performance of radar-based systems is affected by obstacles, signal attenuation, and interference. Therefore, specific environmental requirements must be considered while developing an optimal HAR solution that balances convenience, accuracy, and privacy.

† Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics & Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412

1 M. S., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: sincere96@g.kmou.ac.kr, Tel: 051-410-4822

2 Ph. D. Candidate, Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: seoluck77@gmail.com, Tel: 051-410-4822

3 CEO, Division of Research & Development, CVI, E-mail: ceo@cvi.re.kr, Tel: 032-214-2450

Recently, radar-based HAR technology has been actively researched as a method for automatic learning and classification by applying it to deep learning (DL) models **[5]-[14]**. However, DL models require considerable computational costs and parameters to process and classify high-dimensional radar datasets, making them difficult to commercialize on mobile devices. As the training datasets for radar signals are generally much smaller than those for images, their accuracy suffers significantly. To solve this problem, this study proposes a lightweight radar-based HAR DL model using Self-Attention technology. The proposed model utilizes a combination of a Convolutional Neural Network (CNN) to extract the regional characteristics of the radar signal and a Long Short-Term Memory (LSTM) network to extract its temporal characteristics. Using effectively extracted spatiotemporal information, this study aims to minimize the number of CNN and LSTM layers and improve accuracy using a Self-Attention model that learns the characteristics of mutual relationships.

2. Related works

Advances in DL algorithms have been applied in various fields to implicitly transform low-level input data into high-level data using multiple layers of neural networks. In addition, DL algorithms have been proven to solve difficult classification problems in various fields by learning complex features. Consequently, DL has advanced the development of many fields, including HAR. Kim *et al.* **[5]** first applied CNN to interpret the time-Doppler (TD) map as a 2D image.

CNN, one of the most successful DL algorithms, can reduce the complexity of a network by

reducing the number of weights and replacing the usual matrix multiplication of a typical neural network with a convolution operation. Thus, CNNs can extract local features of a micro-Doppler signatures (MDS) that vary with human activity. Bai *et al.* **[8]** proposed a CNN model consisting of dual input channels that divided the window length of a short-time Fourier transform (STFT) into two parts according to the characteristics of the torso and limbs.

Unlike CNNs that interpret radar data as images, Zhu *et al.* **[9]** proposed a TD map-based Recurrent Neural Network (RNN) for continuous HAR. An RNN comprises three layers: input, hidden, and output. The most important feature is that the hidden layers are interconnected. Therefore, the data from the previous point in each time step can be considered the next data point. As RNNs can mine temporal and semantic information, they can be used to model temporal sequences. With this advantage, the RNN can extract spectral features that change sequentially along the time axis of the MDS. Therefore, RNNs are more robust than CNNs in terms of continuous changes in human behavior.

To reduce the complexity of DL models, methods for optimizing the underlying convolution operations and neural network architecture have been developed. Computationally efficient CNN architectures such as MobileNet **[15]** and ShuffleNet **[16]** have been developed for mobile devices with limited computing resources. However, their application to radar spectrograms is not as efficient as that in computer vision applications. This is because the training dataset for radar signals is small compared to the image dataset, resulting in poor accuracy of well-designed, efficient networks. Therefore, to solve this problem, we attempted to reduce complexity by utilizing Self-Attention in combination with the CNN-LSTM model.

3. Proposed method

3.1 Overview of the proposed method

Figure 1 shows the overall architecture of HAR using radar. The proposed network consists of a continuous wavelet transform (CWT) for data preprocessing, CNN-LSTM layer for feature extraction, and Self-Attention layer for global pattern

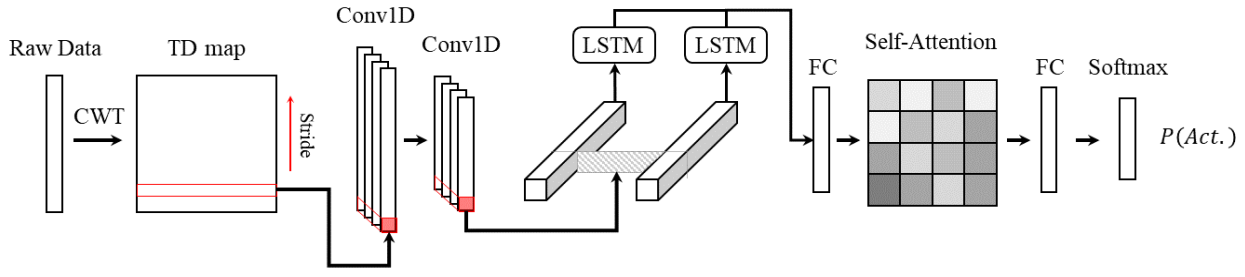

Figure 1: Overall architecture of the proposed method based on the Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) model with Self-Attention

learning. First, raw data are created as a TD map using the CWT, which can adjust the resolution by setting a flexible window size. The TD map was separated into one-dimensional data on the time axis and input into the CNN-LSTM layer, which consisted of a second layer of 1D-CNN for learning regional features and LSTM for learning temporal features. 1D-CNN can preserve temporal information by extracting local features from adjacent times. Because the TD map processed by the 1D-CNN preserves temporal information, it can be viewed as time-series data with multiple channels, and time-series data can be processed using the LSTM. The feature vectors processed in the CNN-LSTM layer reduce unnecessary calculations through the use of the Self-Attention layer. Finally, a classifier was added to obtain the HAR results.

3.2 CNN-LSTM

The proposed CNN-LSTM model uses a 1D-CNN to extract regional features over time from the TD map obtained through CWT and an LSTM to extract the time characteristics. A CWT can increase the resolution in both the time and frequency domains because a window size that varies according to the frequency size can be set when applying a fast Fourier transform (FFT). Therefore, the raw data were converted into a TD map using CWT. To extract the Doppler information each time from this TD map, it must be embedded into one-dimensional data. Therefore, in this study, a 1D-CNN was applied to extract local feature vectors over time from the TD map. The spectrograms were treated as 1-D (time dimension) signals with multiple channels (frequency dimension) that can be analyzed using a 1D-CNN. Consequently, the temporal characteristics of the spectrogram can be preserved and exploited more effectively in LSTMs. The computational complexity of a 1D-CNN is also lower. Unlike 2D-CNN, a 1D-CNN has advantages in analyzing time-series data because the window moves in only one direction while maintaining the local feature information. The TD map was inputted into the 1D-CNN to obtain a feature vector in which the Doppler information of each region was embedded over time.

$$z(m,k) = Conv1D(TD(t_m, f_n), W_k) \quad (1)$$

where $Conv1D$ denotes the convolution operation in the spatial domain, output $z(m,k)$ where k is the number of filters.

In addition, LSTM, which is primarily used for time-series data analysis, was applied to extract the Doppler information according to the time change of the data.

$$h(m,u) = LSTM(z(m,k), h(m-1)) \quad (2)$$

where $h(m,u)$ is the LSTM output at time m, u is the size of the hidden layer, and input $z(m,k)$ is the input at time m. An LSTM is advantageous for time-series data analysis because it uses information from the previous data to predict the next data. As a result, the output of the CNN-LSTM model has all the spatiotemporal characteristics because it includes local Doppler information according to the time change. These data were converted into one-dimensional data to be input into the subsequent Self-Attention model.

3.3 Self-attention and classification

The TD map based on the radar signal has a temporal correlation, according to Doppler. Therefore, because the temporal information was preserved without being destroyed in the previous CNN-LSTM network, the data still had temporal relevance. In other words, the final output feature vector of the CNN-LSTM can be regarded as the embedded Doppler information over time. However, the feature vector combined with the visual axis may contain parts with core patterns. Self-Attention analyzes the correlation between one's own data based on a scaled dot product, which analyzes the correlation among the input sequential data.

$$S(l,e) = softmax(V_{att}\tanh(U_{att}h'(m,u)) \cdot h(m,u) \quad (3)$$

where $h'(m,u)$ is the transpose of $h(m,u)$, $U_{att} \in R^{LA\times U}$ and $V_{att} \in R^{LO\times U}$ are the weight matrices forming the attention module; LA represents the attention length; LO represents the length of the output; and the dot product $\cdot$ is taken in the spatial domain. In general, patterns of human behavior do not exist evenly in all time domains but can be classified through important behaviors at small moments. Existing methods require a large amount of unnecessary computation because they learn patterns in all parts.

In this study, we reduced the overall scale of the model by collecting and calculating only the necessary parts of the data using Self-Attention. **Figure 1** shows the Softmax function for this Self-Attention and classification. P (Act.) refers to the probability based on action. Each feature vector output through the LSTM proceeds with concatenation to insert the input of Self-Attention, and this combined vector is divided into an attention query, key, and value. Because Self-Attention analyzes one's own data, the query, key, and value are all composed of the same data. The result of Self-Attention is a normalized feature vector according to

the importance of the data, which is projected to the final action size through the next fully connected layer. This is a one-hot encoded result and presents the final probability. Unlike existing large models, the proposed model can achieve the same performance with a minimized amount of computation through Self-Attention and the classifier.

4. Experiment and results

4.1 Datasets and training environment

The performance of the proposed method was validated by collecting data from different locations. This is significant because the effect of Doppler generation generally depends on the radar locations. Therefore, the open dataset acquired by the University of Glasgow was used for the analysis, and the data were collected from six different locations using an Ancortek (SDR-580AD) radar. The center frequency and chirp duration were 5.8 GHz and 1 ms, respectively, and the sample rate was 128 beats per sweep (1 ms). Sixty participants from different age groups performed six activities: 1) walking (data size: 213), 2) sitting (222), 3) standing up (220), 4) drinking water (219), 5) picking up an object from the floor (220), and 6) falling (188). All participants repeated each activity three times. Walking was associated with radar data acquired for 10 s, whereas the other actions involved data collected for 5 s. A total of 1282 samples were collected.

In this study, the training and testing experiments were performed using a single NVIDIA GeForce RTX 2080Ti graphics processing unit. All the models were optimized using the Adam optimizer with a learning rate of 0.00001. HAR models were developed using TensorFlow (version 1.15.0, https://www.tensorflow.org/) and Keras (version 2.3.1, https://www.keras.io/). In addition, the training data were equally divided into batches with a batch size of 16 samples, and the neural network was trained for 100 epochs. The final training model was selected as the testing model regardless of the validation loss during the training process using the cross-entropy loss.

4.2 Experimental results

The effectiveness of the proposed radar-based HAR was verified by performing an experiment with respect to location. The data matrix was reshaped to a size of 128 × 10,000, where 128 denotes the number of time samples per sweep and 10000 indicates the number of chirps for the measurement cycle of one activity sample. The FFT was applied to 128 points in the data matrix acquired in one sweep. The FFT results contained 63 range points, each with a resolution of 37.5 cm. For each activity, the number of points was in the range of 5–25 (21 points). In the case of CWT, MATLAB (R2019a, MathWorks, Natick, MA, USA) was used for the mother wavelet, and the scale parameters of the Morse wavelet were $\beta = 40$ and $\gamma = 3$.

We also compared our method with 2D-CNN and CNN-LSTM models. The input of the TD map was resized to 144 (time) × 108 (Doppler) bins. The input of the network used a TD map based on CWT. The 2D-CNN comprised four convolutional layers, followed by two fully connected layers with 32 outputs. The number and size of 2D-CNN kernels were set to 64–128 and 5–3, respectively. Two max-pooling layers were used after every two convolutional layers. Furthermore, the number and size of the 1D-CNN kernels were set to 32–64 and 5–3, respectively. The LSTM was composed of a two-stage bi-LSTM with 128 memory blocks and a dropout rate of 0.25.

The experiment used cross-validation to evaluate the model's performance by considering a DL model. **Table 1** shows the accuracy of each model using 5-fold cross-validation. The total numbers of parameters for the 2D-CNN, 1D-CNN-LSTM, and 1D-CNN-LSTM-Self-Attention models were 2,073,542, 758,214, and 115,622, respectively. Compared to the 2D-CNN model, the proposed model reduces the number of parameters by 17.93 times by using Self-Attention and light weighting the remaining layers. Furthermore, the proposed model showed the highest accuracy in all experiments except for Fold 3, and the average accuracy was shown to improve by 2.03% compared with that of the 1D-CNN-LSTM model.

Table 1: Comparison of different deep learning models for human activity recognition.

Model	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Average (%)
1. 2D-CNN **[5]**	90.27	92.61	90.63	93.36	93.36	92.04
2. 1D- CNN-LSTM **[9]**	92.60	92.61	**94.92**	94.14	91.80	93.21
3. 1D-CNN-LSTM-Self-Attention	**93.77**	**96.50**	93.75	**96.09**	**96.09**	**95.24**
Average (%)	92.21	93.91	93.10	94.53	93.75	-

5. Conclusion

We propose a lightweight HAR DL model using Self-Attention technology to address the complexity and computational costs of DL models. The proposed HAR model using radar signals involves a CNN-LSTM layer for feature extraction and a Self-Attention layer for global pattern learning. The TD map obtained through CWT was used as the input, and a 1D-CNN was applied to extract local feature vectors over time. The LSTM layer was used to extract the Doppler information according to the time change of the data. The final output feature vector of the CNN-LSTM layer includes the embedded Doppler information over time, and Self-Attention is used to analyze the correlation among the input sequential data and reduce the overall scale of the model by collecting and processing only the necessary parts of the data. The proposed method achieves the same performance as existing large models with minimal computation through Self-Attention and classification. Therefore, this lightweight DL model using Self-Attention technology has great potential in various fields, especially mobile device usage, owing to its lower computational complexity and memory capacity.

Acknowledgement

This work was supported by the Technology develoment Program(S3220267) funded by the Minitry of SMEs and Startups(MSS, Korea).

Author Contributions

Conceptualization, H. I. Seo; Methodology, S. B. Jeong; Software, S. B. Jeong and H. I. Seo; Data curation S. B. Jeong and G. D. Choi; Writing-Original Draft Preparation, S. B. Jeong; Writing-Review & Editing, D. H. Seo; Supervision, D. H. Seo.

References

[1] H. F. Nweke, Y. W. Wah, G. Mujtaba, and M. A. Al-garadi, “Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," Information Fusion, vol. 46, pp. 147-170, 2019.

[2] A. B. Mabrouk and E. Zagrouba, “Abnormal behavior recognition for intelligent video surveillance systems: A review,” Expert Systems with Applications, vol. 91, pp. 480-491, 2018.

[3] M. Z. Uddin and A. Soylu, “Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning,” Scientific Reports, vol. 11, 2021.

[4] J. Le Kernec et al., “Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms,” IEEE Signal Processing Magazine, vol. 36, no. 4, pp. 29-41, 2019.

[5] Y. Kim and T. Moon, “Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks,” IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 1, pp. 8-12, 2016.

[6] A. Helen Victoria and G. Maragatham, “Activity recognition of FMCW radar human signatures using tower convolutional neural networks,” Wireless Networks, pp. 1-17, 2021.

[7] F. J. Abdu, Y. Zhang, and Z. Deng, “Activity classification based on feature fusion of FMCW radar human motion micro-Doppler signatures,” IEEE Sensors Journal, vol. 22, no. 9, pp. 8648-8662, 2022.

[8] X. Bai, Y. Hui, L. Wang, and F. Zhou, “Radar-based human gait recognition using dual-channel deep convolutional neural network,” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 12, pp. 9767-9778, 2019.

[9] J. Zhu, H. Chen, and W. Ye, “A hybrid CNN–LSTM network for the classification of human activities based on micro-Doppler radar,” IEEE Access, vol. 8, pp. 24713-24720, 2020.

[10] H. -U. -R. Khalid, A. Gorji, A. Bourdoux, S. Pollin, and H. Sahli, “Multi-view CNN-LSTM architecture for radar-based human activity recognition,” IEEE Access, vol. 10, pp. 24509-24519, 2022.

[11] S. Zhu, R. G. Guendel, A. Yarovoy, and F. Fioranelli, “Continuous human activity recognition with distributed radar sensor networks and CNN–RNN architectures,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-15, 2022.

[12] W. Ding, X. Guo, and G. Wang, “Radar-based human activity recognition using hybrid neural network model with multidomain fusion,” IEEE Transactions on Aerospace and Electronic Systems, vol. 57, no. 5, pp. 2889-2898, 2021.

[13] X. Li, Y. He, F. Fioranelli, X. Jing, A. Yarovoy, and Y. Yang, “Human motion recognition with limited radar

micro-Doppler signatures," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 8, pp. 6586-6599, 2021.

[14] Z. Cao, Z. Li, X. Guo, and G. Wang, "Towards cross-environment human activity recognition based on radar without source data," IEEE Transactions on Vehicular Technology, vol. 70, no. 11, pp. 11843-11854, 2021.

[15] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," pp. 1-9, 2017.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848-6856, 2018.