Dilated Causal Convolution Based Human Activity Recognition using Voxelized Point Cloud Radar Data

Samuel Kakuba, Savina Jassica Colaco, Jung Hwan Kim, Dong Gyu Lee, Young Jin Yoon, Dong Seog Han*
School of Electronics and Electrical Engineering
Kyungpook National University
Daegu, Republic of Korea
2021327392@knu.ac.kr, savinacolaco@knu.ac.kr, jkim267@knu.ac.kr, jasmindoe@knu.ac.kr, skag2603@knu.ac.kr, dshan@knu.ac.kr*

***Abstract*—Due to the immense advantages that include contactless sensing, privacy-preserving, and lighting condition insensitivity, radar systems have been applied in Human Activity Recognition (HAR). The radar signal is often used in its raw form, pre-processed into micro-Doppler signatures or represented as voxelized Point clouds. However, the point cloud data is usually sparse and non-uniform. HAR deep learning models ought to learn the spatial and temporal features. These models should be robust for all considered activities and computationally efficient. Instead of other deep learning techniques used in literature, dilated causal convolutions (DCC) provide a broad receptive field with a few layers while preserving the resolution of the inputs throughout the model, thereby learning the spatial and temporal cues. In this paper, we investigated the use of DCC in combination with other deep learning techniques like residual blocks (RDCC), transformer encoders (TED), and bidirectional long-short-term memory (BiLSTM). We subsequently proposed the DCCB model that consists of DCC layers and BiLSTM layers. The proposed model exhibits a commendable performance in terms of accuracy, and generalization especially in terms of balanced robustness for all activities.**

***Index Terms*—activity recognition, dilated convolutions, radar data**

I. Introduction

Deep learning techniques in human activity recognition (HAR) have led to the improved lifestyle of people especially the elderly through monitoring their daily living activities. Besides the camera and wearable sensors, the radar sensors can be utilized to track and detect human activities. They exhibit advantages of contact-less sensing, lighting condition insensitivity, and privacy preservation compared to their counterparts since they only use the radar signal to sense the activity of the non-stationary target [1]. The radar sensor captures signals that depict actions performed either away or towards its line of sight. The received radar signal depicts and can measure the range, and Doppler information about target objects [2]. The range time map (RTM), Doppler time map (DTM), and angle time map (ATM) can be computed, extracted, and used as inputs for deep learning models [3]. Several HAR studies that use the radar sensor data have been proposed. An adaptive threshold method that highlights the region of interest in micro-Doppler signatures was proposed in [4]. A deep learning model that detects continuous HAR activities was proposed in [5]. These works obtain features extracted from the radar signal through signal processing as input to the deep learning models. Though commendable results are obtained, the computational cost involved makes it expensive for end-to-end models of this nature to be deployed in low-resource devices. The radar signal can also be transformed into three-dimensional point cloud data that depicts the range, velocity, elevation, and azimuth angles of the target. The point cloud can subsequently be used as the input of the deep learning models. The point cloud can be voxelized [6] or transformed into a multi-view representation [7]. The radar sensors produce point clouds that are sparse and non-uniform which makes it hard for the deep learning models to learn the local context that exists in them. The point cloud voxelization method is used in [6] to alleviate this problem. However, voxels increase the memory and computational requirements. As observed in [8], [9] and [10], DCC layers provide a broad receptive field with a few layers while preserving the resolution of the inputs throughout the model. We therefore investigate the use of DCC in combination with other deep learning techniques like RDCC, TED, and BiLSTM to solve the computational efficiency problem and improve HAR model robustness. We subsequently propose the DCCB model that consists of DCC and BiLSTM layers that is robust and computationally efficient according to our experiments.

The rest of the paper is organized as follows: the methods are presented in Section II. Section III presents the results and discussion. The paper is concluded in Section IV.

II. Methods

In this section we present the datasets, the proposed model and the experiments carried out in this paper.

Dataset: We used the mmActivity radar point cloud dataset that was provided in [6]. The point cloud datasets are used to create their voxelized representation which are fed as inputs into the models. In our experiments, we evaluated different dimensions of the voxelized representations of point clouds together with their velocity to find out the behavior of the

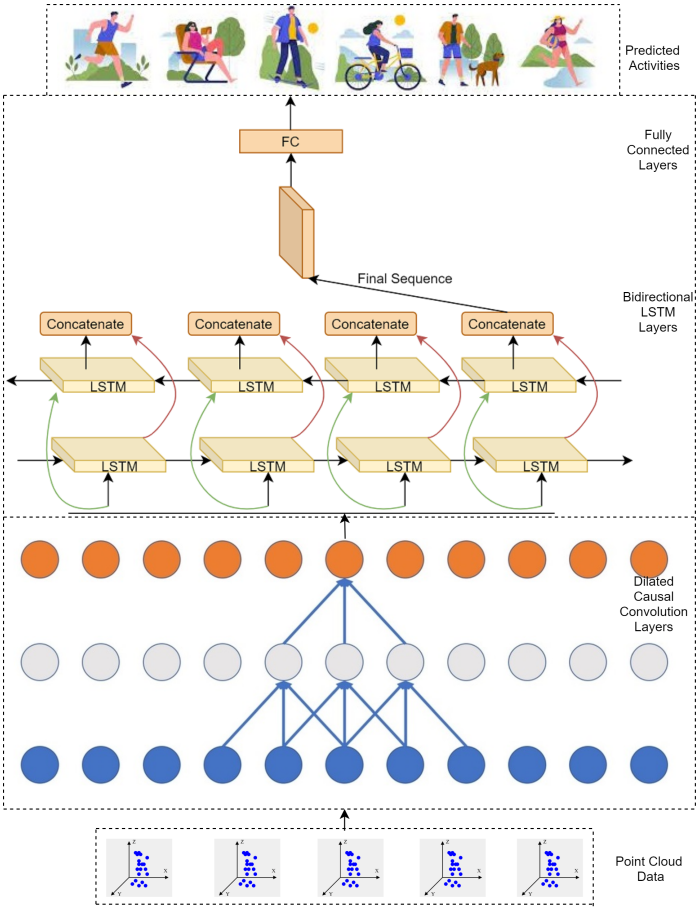

Fig. 1: The proposed DCCB model that consists of dilated causal convolution (DCC) and BiLSTM layers.

models and the performance they exhibit. Because of memory constraints and the need for computationally efficient models, we used a voxel representation of input dimensions (60*10*8*8).

The Proposed Model: We propose the DCCB model that consists of the DCC and BiLSTM layers. This model ensures robust and computationally efficient activity recognition. Though the BiLSTM alone or in combination with the time distributed CNN used in [6] exhibits a commendable performance, it fails to be uniformly robust on all the activities in addition to being computationally expensive. It specifically confuses activities that involve a similar change in velocity like jumping, jumping jacks, and walking. In addition, the BiLSTM encounters problems of slow convergence and sluggish training that uses a lot of memory resources and encounters the vanishing gradient problem. Like the traditional CNN, the time-distributed CNN does not capture temporal clues over time which are crucial for continuous HAR. In addition, its receptive field is limited to the kernel size used making its selection crucial for the model's efficiency. A very small kernel may miss global patterns yet a too large one may overlook local information. In this paper, we used dilated causal convolution layers in combination with BiLSTM layers as shown in Fig. 1. The dilated causal convolutions help the model to use a large receptive field which is crucial for effective feature extraction with less increase in the number of parameters compared to the number of layers and consider spatial as well as temporal cues among the extracted features. The BiLSTM handles long-term dependencies.

Experiments: To ascertain the combination of techniques that can exhibit the most robust results in terms of generalization, confusion ratio, and computational efficiency, we carried out some experiments. They included models with only DCC layers, only RDCC blocks, a combination of DCC and TED layers, a combination of RDCC blocks and TED layers, a combination of RDCC blocks, TED and BiLSTM layers, a combination of RDCC and BiLSTM layers, and a combination of DCC and BiLSTM layers which forms the DCCB model we present in this paper.

III. Results and Discussion

A. Results

The results shown in Table I show the performance of different combinations of deep learning techniques with DCC layers. The results are reported in terms of accuracy (A), loss (L), and confusion ratio of jumping (CRJ) which is often confused with jumping jerks and walking by models. We also present confusion ratio (CRB) results for boxing. To assess the computational complexity, we present the total number of parameters for each of the models. The confusion matrices for two particular experiments which show the benefit of Dilated convolution but also help us explore the trade-off between computational complexity and robustness are shown in Fig. 2. The performance of the proposed DCCB model compared to the other models investigated in this paper shows that DCC layers are significant in the performance of the models that use sparse and nonuniform point cloud radar data however a trade-off between complexity and robustness needs to be given careful attention.

B. Discussion

As shown in Table I and the confusion matrices in Fig. 2, the DCC operation gives a better receptive field for the sequential data in addition to learning the long-term dependencies. This is why each model gives a good accuracy. However, besides the accuracy, it is important to assess the robustness of the model for all the activities in the dataset and the computational complexity of the model. From the results, we observe that models that utilize the RDCC blocks have less computational complexity compared to the models that use DCC layers. The situation is similar even when a transformer encoder that utilizes multi-head attention of 4 to 8 heads is used in combination with the RDCC blocks. This is because the residual nature caused by the gated activation units (GAU) [8] and concatenation does not increase the number of parameters

TABLE I: Performance of the different models used in our experiments.

Model	A(%)	L	CRJ(%)	CRB(%)	Parameters
DCC	80.87	0.662	77.00	85.00	6,850,565
RDCC	87.16	1.274	70.00	81.00	2,264,325
DCC-TED	87.45	0.943	69.00	85.00	7,437,861
RDCC-TED	**79.23**	1.692	**59.00**	95.00	**1,834,277**
RDCC-TED-BiLSTM	85.53	1.072	78.00	70.00	1,965,861
Proposed DCCB	**90.01**	1.055	81.00	85.00	**12,055,045**

(a) (b)

Fig. 2: The confusion matrix results. (a) The RDCC-TED Model. (b) The proposed DCCB Model.

while providing a large receptive field. However, the benefit of the RDCC block is more suitable for purely sequential data like text and audio but less impactful for voxelized point cloud data. Though the number of parameters increases for models with DCC layers, their performance in terms of generalization is incredible for voxelized point cloud data. This is why the models that utilize them have a better accuracy and confusion ratio for often confused activities. The proposed DCCB model considers the long-term dependencies by using the BiLSTM layers as well as solving the vanishing gradient problem which contributes to its better generalization and robustness.

Prediction Error Analysis: As observed in confusion matrices shown in Fig. 2 (a) and 2 (b), the individual class accuracy predicted by the two models we have selected shows the significance of dilated convolutions. As observed in Fig. 2 (a) the RDCC-TED model which has the lowest parameters and accuracy as reported in Table I has the worst confusion ratio for jumping and Jumping jacks. A similar reasoning can be applied to the other models whose confusion matrices have not been shown because of lack of space.

Generally, though the proposed DCCB model exhibits a comparatively similar accuracy with the RadHAR [6] that proposed the use of BiLSTM and time-distributed CNN, its generalizability and therefore balanced robustness for all activities is commendable.

IV. Conclusion

In this paper, we investigated the use of dilated causal convolutions in combination with other deep learning techniques like residual blocks, transformer encoders (TED), and bidirectional long-short-term memory (BiLSTM) to solve the robustness and computational efficiency problem. We subsequently proposed the DCCB model that consists of dilated causal convolution (DCC) layers and BiLSTM layers. The proposed DCCB model is robust for all considered activities and computationally efficient according to our experiments. However, it is worth exploring multi-view point cloud representations for activity recognition.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A1A03043144).

REFERENCES

[1] G. Bhavanasi, L. Werthen-Brabants, T. Dhaene, and I. Couckuyt, "Patient activity recognition using radar sensors and machine learning," *Neural Computing and Applications*, vol. 34, no. 18, pp. 16 033–16 048, 2022.

[2] L. Zheng, J. Bai, X. Zhu, L. Huang, C. Shan, Q. Wu, and L. Zhang, "Dynamic hand gesture recognition in in-vehicle environment based on fmcw radar and transformer," *Sensors*, vol. 21, no. 19, p. 6368, 2021.

[3] S. Ahmed, W. Kim, J. Park, and S. H. Cho, "Radar-based air-writing gesture recognition using a novel multistream cnn approach," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 869–23 880, 2022.

[4] Z. Li, J. Le Kernec, Q. Abbasi, F. Fioranelli, S. Yang, and O. Romain, "Radar-based human activity recognition with adaptive thresholding towards resource constrained platforms," *Scientific Reports*, vol. 13, no. 1, p. 3473, 2023.

[5] 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.

[6] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 51–56.

[7] F. Luo, S. Khan, A. Li, Y. Huang, and K. Wu, "Edgeactnet: Edge intelligence-enabled human activity recognition using radar point cloud," *IEEE Transactions on Mobile Computing*, 2023.

[8] S. K. Pandey, H. S. Shekhawat, and S. Prasanna, "Emotion recognition from raw speech using wavenet," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 1292–1297.

[9] S. Kakuba and D. S. Han, "Speech emotion recognition using context-aware dilated convolution network," in *2022 27th Asia Pacific Conference on Communications (APCC)*. IEEE, 2022, pp. 601–604.

[10] S. Kakuba, A. Poulose, and D. S. Han, "Attention-based multi-learning approach for speech emotion recognition with dilated convolution," *IEEE Access*, vol. 10, pp. 122 302–122 313, 2022.