Optimized Data Transmission for Radar-Based Edge-Cloud Human Activity Recognition via Quantization

Victor Tuytte, Arash Heidari, Lorin Werthen-Brabants, Tom Dhaene, Ivo Couckuyt
IDLab, Ghent University - imec, Belgium
{victor.tuytte, arash.heidari, lorin.werthenbrabants, tom.dhaene, ivo.couckuyt}@ugent.be

***Abstract* — Radar systems offer a privacy-preserving approach to human activity recognition, leveraging Deep Neural Networks (DNNs) for effective data processing. However, the scalability of large DNNs becomes challenging when handling multiple radar streams simultaneously, such as monitoring several rooms in a hospital. A viable solution, the Split Bidirectional Recurrent Neural Network (Split BiRNN) for Human Activity Recognition, addresses this issue by dividing data processing between on-premise (low-power, low-cost devices) and off-premise (higher performance devices) components. This paper delves into the optimization of data transmission between these two components, considering the significant battery consumption of edge devices due to data transmission. Through quantization techniques, the paper explores the reduction of data size while maintaining accuracy levels. Results indicate that quantization to 2 or 1 bits is feasible, with minimal accuracy loss of 0.21% and 1.12%, respectively.**

***Keywords* — Radar, Human Activity Recognition, Bidirectional RNN, Edge-Cloud Interaction, Quantization, Binary Quantization**

I. Introduction

Real-time human activity recognition within hospital rooms or nursing homes holds significant importance, particularly in swiftly detecting critical events such as a patient falling. This urgency is particularly pronounced among geriatric patients [1], [2], who are more prone to enduring severe injuries from falls, especially if timely intervention is lacking. Leveraging radar technology offers a privacy-preserving means to detect falls promptly. Deep Learning (DL) emerges as a popular technique for fall detection or human activity recognition, albeit entailing prolonged training times and demanding computational resources during prediction. As DL is deployed at scale to manage numerous radar streams simultaneously, the associated costs escalate significantly.

To address these challenges, Split BiRNNs [3] have been introduced. This approach tackles the demands of handling multiple radar streams by employing a two-staged model. Designed to operate across separate devices, each handling distinct Micro-Doppler [4] streams, this framework enables real-time monitoring of human activities across diverse environments, such as hospitals with multiple rooms. The two-stage design unfolds as follows:

1) An edge device executes lightweight models to compute class predictions on incoming radar frames, facilitating real-time streaming of results to parties necessitating immediate feedback.
2) A more capable device, equipped to handle substantial data processing in batches, applies a backward model on intermediate computations performed by the edge devices, refining predictions. This higher-capacity device rectifies any inaccuracies in predictions made by the edge device.

This paper concentrates on optimizing the communication bandwidth between the lightweight model deployed on the edge device and the more capable backward model, thereby enhancing the scalability and efficiency of the Split BiRNN, as data transmission stands out as a large source of power consumption for edge devices [5], [6]. The approach involves analyzing and quantizing the intermediate computations transmitted from the lightweight model to the backward model. Quantization of these intermediate computations significantly reduces the amount of data exchanged between the models, thereby minimizing the communication bandwidth required. Through the implementation of this optimization, the burden of data transmission for edge devices is alleviated. Previous research in human activity recognition demonstrates the feasibility of creating fully binarized DNNs or 2-bit DNNs with only a marginal loss in accuracy [7], [8]. However, the previous research is not radar-based; instead, it utilizes a camera for human activity recognition.

II. Quantization of deep neural networks

Quantization is one among several optimization techniques available for optimizing DNNs. Its primary objective is to decrease the precision of the model's weights and/or activations [9]. This process encompasses both floating point quantization and integer quantization. Floating point quantization reduces the bit representation of floating point numbers, e.g., from 32 bits to 16 bits, while integer quantization maps values to fixed-point representation, requiring a specified range for quantization.

Quantization broadly falls into two categories: Post-Training Quantization (PTQ) and Quantization Aware Training (QAT) [9], [10]. PTQ involves applying quantization after the model has undergone training, such as quantizing the weights subsequent to training the model with full precision. This method proves cost-effective as it removes the need for retraining, allowing the original weights/activations to be reused for computing the new quantized versions. However, PTQ requires knowledge of the value range to be quantized. While for

weights, it is feasible to determine the minimum and maximum values for quantization, for activations, it necessitates running several training samples to ascertain the value range accurately. However, PTQ may lead to a noticeable decline in accuracy since it does not involve retraining any component of the model [9].

Conversely, QAT incorporates model retraining. Its core concept revolves around making the model conscious of quantization during training. This is achieved by simulating the quantization process during the forward pass to compute the model's loss, utilizing only actual lower-bit weights/activations during inference. This approach enables the model to compensate for the minor errors induced by quantization. Nevertheless, estimating the range of values for activations becomes more challenging with QAT. According to [11], employing an exponential moving average proves beneficial in estimating the range of activation values.

A radical form of quantization is binary quantization, which limits weight/activation values to just 1 bit [12]. Typically, a prevalent strategy involves representing the values -1 and 1 with binary values 0 and 1, respectively. This conversion is straightforwardly accomplished by employing the sign function. Nonetheless, binary quantization frequently leads to a significant reduction in model accuracy for complex models [9].

The primary advantages of quantization lie in reducing the storage demands of DNN models and decreasing the memory bandwidth necessary during model inference. Specifically, for this paper, the ability to decrease the bit depth of the activation layer between the forward and backward branches of the split BiRNN model is interesting for minimizing the volume of data that needs to be sent from the forward branch on the edge device to the backward branch on the cloud.

III. Experiments

A. Dataset and Model

a) Dataset: The model is trained and evaluated using the PARRad dataset [3], [13]. This dataset comprises 22 hours of radar data, which is divided into two subsets: Homelab and Hospital, both designed to simulate hospital rooms. This study specifically concentrates on the Hospital subset within PARRad, encompassing 13,359 activities across 9 distinct classes. The Hospital dataset encompasses observations from 20 test subjects engaged in various activities across four separate 10-minute sessions.

The dataset features Micro-Doppler (MD) signatures captured using Texas Instruments (TI) xWR14xx and TI xWR68xx radars, operating at center frequencies of 77 GHz and 60 GHz, respectively. Our analysis leverages MD signatures over time, as outlined in previous research [4], which encode speed relative to the radar sensor. These signatures are obtained from different corners of the hospital room, capturing activities concurrently. Consequently, each capture session is effectively duplicated, although variations exist in the field of view of the radar sensors.

Each captured MD signature comprises 128 Doppler bins, derived from averaging data from 93 range bins within corresponding RD maps. These MD signatures are temporally stacked, resulting in a datapoint $\mathbf{X} \in \mathbb{R}^{t \times 128}$ for each recording in the dataset, where t varies across recordings. Typically, t approximates t = 6666, corresponding to 10 minutes of real-time due to the 0.09-second frame length [14].

b) Model: The model is adopted from [3], which is a hybrid CNN-GRU approach. It comprises of a CNN that supplies inputs to both forward and backward branches. Specifically, the forward branch comprises of three sequentially connected GRUs followed by a fully connected layer. Similarly, the backward branch also comprises of three sequentially connected GRUs followed by a fully connected layer, with each GRU also concatenating their output with the output of the corresponding GRU in the forward branch. The data transmission discussed in this paper involves the output of the CNN sent to the backward branch and the intermediate results of the GRUs from the forward branch transmitted to the backward branch. This is visualized in Figure 1.

Notably, unlike the approach in [3], the fully connected layer is not shared between the forward and backward branches, ensuring their complete separation. Another notable change for the model used in this paper compared to [3] is the alteration of the activation function at the output of the CNN from a relu function to a $\tanh$ function. Despite these changes, the accuracy of the model was minimally affected: the original model achieved backward and forward accuracies of 90.43% and 82.49%, respectively, while the modified model achieved 90.54% and 82.53% for backward and forward accuracies, respectively.

B. Quantization Applied to Split BiRNN

By setting the activation function of the CNN output to $\tanh$, all intermediate computations, including the outputs of the GRUs, which already utilize the $\tanh$ function, are guaranteed to fall within the range of the $\tanh$ function. This ensures that all intermediate computations x now fall within the range of -1 to 1, which simplifies determining the appropriate range for quantifying the intermediate computations. The following formulas were employed to map the values x to their quantized counterparts x' based on the number of bits allocated for a single value, where B represents the width of each quantization step.

$$B = \frac{2}{2^{\text{bits}} - 1} \tag{1}$$

$$x' = \text{round}\left(\frac{x}{B} + \frac{1}{2}\right) \times B - \frac{B}{2} \tag{2}$$

Upon deeper examination of the activation value distribution, it becomes evident that a significant portion of values tends to gravitate towards the extremes of the $[-1, 1]$ range, as illustrated in Figure 2. This observation suggests that binary quantization might be a feasible approach specifically for this activation layer, given that a considerable number of values are already

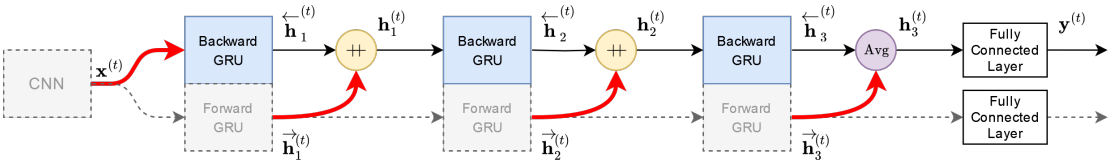


Fig. 1. Data transmission from the forward branch to the backward branch indicated in red. Adapted from [3].

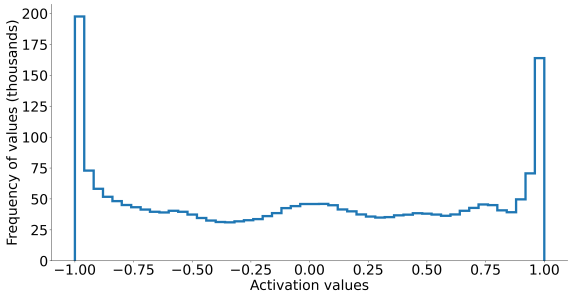


Fig. 2. Histogram of activation values of the third GRU of the forward branch of the Split BiRNN.

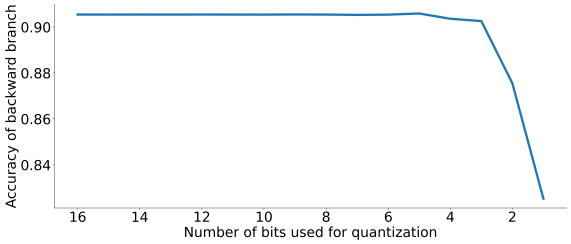


Fig. 3. Backward branch accuracies in function of bits used for PTQ.

concentrated around -1 or 1. Note that by using bits $= 1$, Equations 1 and 2 results in the same behavior as the sign function, mapping all values to either -1 or 1, which corresponds to binary quantization.

C. Communication Bandwidth Reduction

Quantization is applied to the intermediate computations of the forward branch that are being transmitted to the backward branch. This means that the model remains untouched except for the connections between the forward and backward branches, which undergoes quantization. The quantization is performed using formulas 1 and 2 in a PTQ approach, meaning without retraining any weights. This straightforward approach already achieves quantization to small bit depth without a significant decrease in accuracy, for > 4 bits the accuracy deviates less than 0.1% from the non-quantized model. However, it becomes evident that a small decrease in accuracy occurs for 4-bit and 3-bit quantization and a more noticeable decrease in accuracy occurs with 2-bit and 1-bit quantization. This observation is illustrated in Figure 3. This also suggests that employing a pure PTQ approach for binary quantization does not yield practical outcomes, considering that the accuracy of 2-bit quantization at 87.55% still surpasses that of the forward branch at 82.78%, while the accuracy of 1-bit quantization at 82.53% shows only a negligible difference from the accuracy of the forward branch.

The decline in accuracy for fewer bits can be remedied by retraining the model using a QAT approach. This involves reusing the original weights and freezing the CNN and forward branch to prevent them from being retrained, focusing solely on retraining the backward branch while it is aware of the quantization. This does not significantly increase the accuracy for bit quantization where more than 4 bits are used, but it does increase the accuracies for 1-bit, 2-bit, 3-bit, and 4-bit quantization as seen in Table 1.

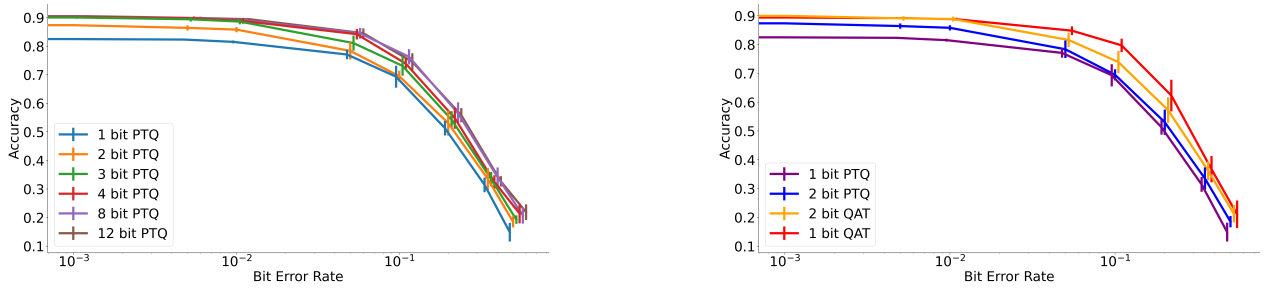
Table 1. Comparison of accuracy scores for PTQ and QAT for 1, 2, 3, and 4 bits.

Bits used for quantization	Quantization techniques	
	PTQ	QAT
4	90.36%	90.46%
3	90.25%	90.48%
2	87.55%	90.32%
1	82.53%	89.41%

These findings indicate that up to 3-bit quantization can be implemented with minimal loss of accuracy for the backward branch, which maintains an accuracy of 90.53% when using non-quantized intermediate computations. Moreover, the decrease in accuracy with 2-bit quantization is merely 0.21%, a statistically negligible figure. However, 1-bit quantization experiences a slight decrease in accuracy of approximately 1.12%, which, in the context of reducing the communication bandwidth between the forward and backward branches by 96.875%, can be considered quite favorable. Note that the reduction of $\frac{31}{32} = 96.875\%$ in communication bandwidth between the forward and backward branches stems from the decrease from 32 bits to 1 bit per value in the quantization process.

D. Error sensitivity

The reduction in bits may render the model more susceptible to bit errors, especially considering the unreliable communication channels typical of edge devices. To address this concern, bit errors were simulated across different quantization levels. This simulation consists of applying bit errors to the quantized values according to a certain rate called the bit error rate (BER). This rate represents how many errors occur on average, e.g. a BER of 10% implies that each bit has a 10% chance of flipping. The model was then evaluated on the test data set for different BERs and quantization levels. Furthermore each combination of BER and quantization level



(a) Comparison of quantization levels, ranging from 1 bit to 12 bit.

(b) QAT vs. PQT.

Fig. 4. Bit Error Rate vs. Accuracy for a different quantization levels and b different quantization methods. Variance markers are slightly spaced apart for readability.

was evaluated multiple times on the test dataset. Conducting multiple evaluations helps average out the randomness of the BER and provides insights into the variance of the accuracy for a given BER and quantization level. The results of this error simulation for the PTQ approach are depicted in Figure 4a. It is evident that 1-bit and 2-bit quantization are slightly more susceptible to bit errors. Lower bit quantization is more susceptible to bit errors in general. Furthermore retraining has negligible influence on the error sensitivity as seen in Figure 4b.

IV. Conclusion and Future Work

In this study, we propose quantizing the intermediate computations exchanged between the forward and backward branches of a Split BiRNN for human activity recognition using radar. Our results demonstrate that employing a Post-Training Quantization (PTQ) approach, i.e. without retraining the model, achieves a negligible loss of accuracy, approximately 0.28%, with 3-bit quantization. Conversely, employing a Quantization Aware Training (QAT) approach, i.e. with retraining the model, yields a similar negligible loss of accuracy, around 0.21%, with 2-bit quantization. Additionally, binary quantization (1-bit quantization) is made feasible by retraining, resulting in a loss of accuracy of approximately 1.12%, which represents a valid tradeoff for reducing communication bandwidth by 96.875%.

The results indicate the feasibility of significantly decreasing communication bandwidth for Split BiRNNs. This capability is particularly advantageous for large-scale deployment of the model and for reducing power consumption in edge devices, where a substantial portion of energy usage is due to data transmission [5], [6].

For future investigations, exploring different optimization techniques such as pruning [9], [10] may further reduce communication bandwidth. Moreover, leveraging more advanced PTQ and QAT techniques could potentially yield improved results.

References

[1] G. F. Fuller, "Falls in the Elderly," *American Family Physician*, vol. 61, no. 7, p. 2159, Apr. 2000, ISSN: 0002-838X, 1532-0650.

[2] X. Yu, "Approaches and principles of fall detection for elderly and patient," in *HealthCom 2008 - 10th International Conference on e-Health Networking, Applications and Services*, Jul. 2008, pp. 42–47. DOI: 10.1109/HEALTH.2008.4600107.

[3] L. Werthen-Brabants, G. Bhavanasi, I. Couckuyt, T. Dhaene, and D. Deschrijver, "Split birnn for real-time activity recognition using radar and deep learning," *Scientific Reports*, vol. 12, no. 1, p. 7436, 2022.

[4] V. C. Chen, F. Li, S. S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: Phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 2–21, Jan. 2006, ISSN: 0018-9251. DOI: 10.1109/TAES.2006.1603402.

[5] S. Al-Sarawi, M. Anbar, K. Alieyan, and M. Alzubaidi, "Internet of things (iot) communication protocols: Review," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 685–690. DOI: 10.1109/ICITECH.2017.8079928.

[6] M. Mahmoud and A. Mohamad, "A study of efficient power consumption wireless communication techniques/ modules for internet of things (IoT) applications," *Advances in Internet of Things*, vol. 6, pp. 19–29, 2016. DOI: 10.4236/ait.2016.62002.

[7] M. Edel and E. Köppe, "Binarized-blstm-rnn based human activity recognition," in *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2016, pp. 1–7. DOI: 10.1109/IPIN.2016.7743581.

[8] Z. Yang, O. I. Raymond, C. Zhang, Y. Wan, and J. Long, "Dfternet: Towards 2-bit dynamic fusion networks for accurate human activity recognition," *IEEE Access*, vol. 6, pp. 56 750–56 764, 2018. DOI: 10.1109/ACCESS.2018.2873315.

[9] F. Daghero, D. J. Pagliari, and M. Poncino, "Chapter eight - energy-efficient deep learning inference on edge devices," in *Hardware Accelerator Systems for Artificial Intelligence and Machine Learning*, ser. Advances in Computers, S. Kim and G. C. Deka, Eds., vol. 122, Elsevier, 2021, pp. 247–301. DOI: https://doi.org/10.1016/bs.adcom.2020.07.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065245820300553.

[10] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.07.045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221010894.

[11] B. Jacob, S. Kligys, B. Chen, *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713. DOI: 10.1109/CVPR.2018.00286.

[12] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/3e15cc11f979ed25912dff5b0669f2cd-Paper.pdf.

[13] 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.

[14] L. Werthen-Brabants, G. Bhavanasi, I. Couckuyt, T. Dhaene, and D. Deschrijver, "Quantifying uncertainty in real time with split birnn for radar human activity recognition," in *2022 19th European Radar Conference (EuRAD)*, 2022, pp. 173–176. DOI: 10.23919/EuRAD54643.2022.9924932.