Decentralised Biomedical Signal Classification using Early Exits

Li Xiaolin[1], *Student Member, IEEE*, Hans Vandierendonck[2], *Senior Member, IEEE*, Dimitrios S. Nikolopoulos[3], *Senior Member, IEEE*, Bo Ji[3], *Senior Member, IEEE*, Barry Cardiff[1], *Senior Member, IEEE*, and Deepu John[1], *Senior Member, IEEE*

2023 21st IEEE Interregional NEWCAS Conference (NEWCAS) | 979-8-3503-0024-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/NEWCAS57931.2023.10198098

***Abstract*—This paper presents a decentralised signal classification approach for data acquired using Internet of Things (IoT) wearable sensors. Traditionally, data from IoT sensors are processed in a centralised fashion, and in a single node. This approach has several limitations, such as high energy consumption on the edge sensor, longer response times, etc. We present a distributed processing approach for convolutional neural network (CNN) based classifiers where a single CNN model can be split into multiple sub-networks using early exits. To reduce the transfer of large feature maps between sub-networks, we introduced an encoder-decoder pair at the exit points. Processing of inputs that can be classified with high confidence at an exit point will be terminated early, without needing to traverse the entire network. The initial sub-networks can be deployed on the edge to reduce sensor energy consumption and overall complexity. We also experimented with multiple exit point locations and show that the point of exit can be adjusted for trade-offs between complexity and performance. The proposed system can achieve a sensitivity of 98.45% and an accuracy of 97.55% for electrocardiogram (ECG) classification and save 60% of the data transmitted wirelessly while reducing 38.45% of the complexity.**

***Index Terms*—ECG classification, Arrhythmia, Decentralised Inferencing, Distributed Network, Deep Learning**

I. Introduction

With the rapid advancement of wearable biomedical sensing, it has become possible to collect long hours of continuous biosignal data, such as ECG. Traditionally, data acquired by a wearable device is first transmitted to an intermediate gateway and then sent to a cloud server for analysis. The automated data processing algorithms are deployed as cloud-native applications for continuous and long-term analysis of ECG data. CNN has shown promising results in ECG signal analysis [1], [2]. However, cloud-deployed systems are highly dependent on communication networks, resulting in high latency and response times, and also cause large energy consumption in wearable devices due to continuous wireless transmission. The straightforward alternative, i.e. to deploy the models fully on edge wearable sensors also faces several challenges. The computational costs of the models can be prohibitively high for edge deployments. The challenge of complexity is inherent to deep learning methodologies, and therefore model compression techniques have been utilized as a means of addressing this issue [3]. Some works also report the development of lightweight machine learning models for edge devices [4]. However, fully edge-native approaches usually lower the system accuracy due to quantization losses or other trade-offs.

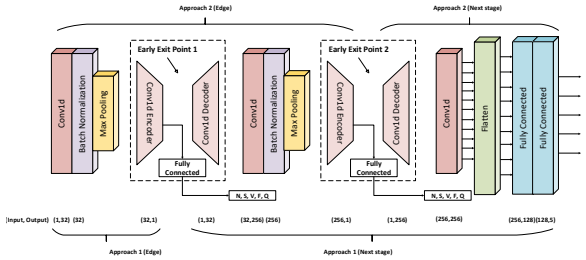

Fig. 1. Decentralised inferencing using 1D CNN with early exit points.

We propose a decentralised data processing approach, where a single CNN model can be split into multiple sub-networks using early exits [5]. Here, each sub-network can accurately classify a subset of the original input and can be deployed onto separate nodes. The cumulative performance of the model (that includes all sub-networks) remains high, similar to that of a large model. The initial sub-networks have lower complexity and therefore can better fit into a resource-limited edge device, making it suitable for real-time applications. Since most inputs are classified by the initial sub-networks, their processing can be terminated early without traversing the rest of the network. This reduces the energy consumption of the edge sensor, saves network bandwidths, and reduces overall latency, response times, and complexity.

II. Architecture

In this work, we used the 1D-CNN originally proposed in [1] as a baseline and added an early exit at one of the two locations in the model (as in Fig. 1). We have optimized the complexity of [1] by changing the first layer's stride, kernel size, and second layer's stride to 5, 25, and 2 respectively while achieving minimal change in performance. All layers before the early exit are used for initial classification, and the remaining layers will be used only for those inputs which didn't exit early. We used ECG from the MIT-BIH Arrhythmia database and mapped them into 5 classes as per AAMI standard [3]. At the early exit, we compute the probability for all 5 classes ($p_i,\ i \in$ (N, S, V, F, Q)), and assign the class with the highest probability ($\max(p_i)$) as the possible output. Further,

[1]University College Dublin, [2]Queen's University Belfast, [3]Virginia Tech. Emails:xiaolin.li@ucdconnect.ie, h.vandierendonck@qub.ac.uk, dsn@vt.edu, boji@vt.edu, {barry.cardiff, deepu.john}@ucd.ie

This work is supported in part by 1) China Scholarship Council 2) Horizon 2020 FET Chist-Era Program and 3) Microelectronic Circuit Centre Ireland.

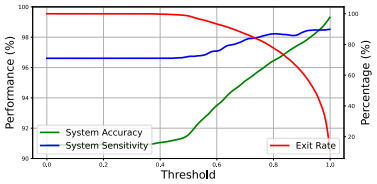

Fig. 2. The system performance and exit rate with early exit point 1.

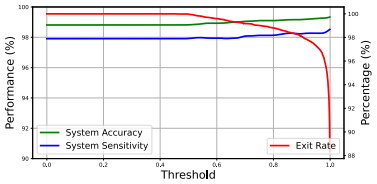

Fig. 3. The system performance and exit rate with early exit point 2.

$\max(p_i)$ is compared against a threshold, and if it is higher, the computation is exited early assuming the initial results are accurate. If $\max(p_i)$ is below the threshold, the classification result is deemed unreliable and processing continues to the next stage of the model. One major issue with this approach is the size of feature maps that needs to be sent from the first to the second sub-network. If each sub-network is deployed on independent nodes (say edge and cloud), it is essential to keep the size of these feature maps minimal, to reduce the communication and energy costs. To address this, we propose to add an encoder-decoder network at the early exit point. The encoder will compress the feature maps, to minimize energy expenditure and communication costs when transmitting data to the cloud. Fig. 1 illustrates the proposed model with exit points for ECG beat classification. The proposed framework facilitates the deployment of the first sub-network onto the edge device, while the remaining layers are deployed on the next node for handling more complex inputs. Only those inputs which can't be classified reliably on the edge are forwarded to the next stage resulting in significant savings compared to the traditional approach, which indiscriminately transmits all inputs to the cloud.

III. Results and Discussion

The MIT-BIH Arrhythmia database is used in this work, with 70%, 15%, and 15% for training, validation, and testing. The performance of the two exit points was evaluated independently in the experiments. Fig. 2 and Fig. 3 illustrate the variations in average system accuracy and sensitivity [3] with respect to increasing thresholds, as well as the percentage of data that can be exited early when either exit point is employed in isolation. The exit rate represents the fraction of the entire dataset that early exits at the specified exit point. The models were optimized for high sensitivity to reduce the chances of not detecting any anomalous classes.

Fig. 2 shows a threshold value of approximately 0.9 achieving a relatively high sensitivity of 98.45% with an accuracy of 97.55% when configuring the first exit point only, enabling early exit for approximately 60% of the data. As Fig. 3 shows, when the threshold is set to 0.97, 97.14% of the data is early exited with 98.26% sensitivity and 99.26% accuracy. Thus, a delayed exit point helps improve accuracy but at the expense of complexity. Table. I displays the performance, the ratio of data that exits early, the number of floating-point operations (FLOPs) at various configurations, and the complexity reduction at the two exit points. The average total FLOPs is calculated by Exit Rate $\times$ FLOPs(Exit) + $(1 -$ Exit Rate$)\times$ FLOPs(Original). Based on the results presented in Table. I, it can be inferred that as the threshold increases, the Exit Rate decreases, despite the potential for performance improvements. Thus, a trade-off must be made between performance and overall complexity.

TABLE I
The exit rate, performance, the number of FLOPs, and reduction percentage at different configurations.

	Threshold	Exit Rate	Accuracy	Sensitivity	# FLOPs	Reduction Percentage
Original Exit Point	-	-	99.32%	98.52%	218880	-
Early Exit Point 1	0.6	94%	93.55%	97%	87040	60.23%
	0.8	77%	96.45%	98.18%	110883	49.34%
	0.9	60%	97.55%	98.45%	134727	38.45%
Early Exit Point 2	0.6	99.6%	98.94%	97.95%	147850	32.45%
	0.8	98.85%	99.1%	98.14%	148385	32.21%
	0.97	97.14%	99.26%	98.26%	149604	31.65%

IV. Conclusions

This work presented a decentralised signal classification approach for biomedical data in wearable sensing applications. We demonstrated that large CNN models can be split into multiple sub-networks using early exits such that each sub-network can be deployed onto independent nodes to address challenges in edge-only or cloud-only deployment of data processing. We also demonstrated that by moving the early exit point to a different location in the model, trade-offs between model performance, resource utilization, and complexity can be achieved. Our proposed model achieved 97.55% overall accuracy and 98.45% sensitivity at the first exit location, reducing 60% volume in wireless transmission and 38.45% arithmetic operations.

References

[1] L. Xiaolin, B. Cardiff, and D. John, "A 1d convolutional neural network for heartbeat classification from single lead ecg," in *2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. IEEE, 2020, pp. 1–2.

[2] G. Sivapalan, K. K. Nundy, S. Dev, B. Cardiff, and D. John, "Annet: A lightweight neural network for ecg anomaly detection in iot edge sensors," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 1, pp. 24–35, 2022.

[3] L. Xiaolin, R. C. Panicker, B. Cardiff, and D. John, "Multistage pruning of cnn based ecg classifiers for edge devices," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1965–1968.

[4] R. N and V. Meshram, "Arrhythmia detection based on hybrid features of t-wave in electrocardiogram," *International Journal of Intelligent Engineering and Systems*, vol. 11, 02 2018.

[5] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2464–2469.