Live Demonstration: Real-time audio and visual inference on the RAMAN TinyML accelerator

Adithya Krishna[1,2*], Ashwin Rajesh[1*], Hitesh Pavan Oleti[1], Anand Chauhan[1] , Shankaranarayanan H[1], André van Schaik[2], Mahesh Mehendale[1] and Chetan Singh Thakur[1]

[1]Department of Electronic Systems Engineering, Indian Institute of Science, Bangalore, India
[2]International Centre for Neuromorphic Systems, The MARCS Institute, Western Sydney University, Australia

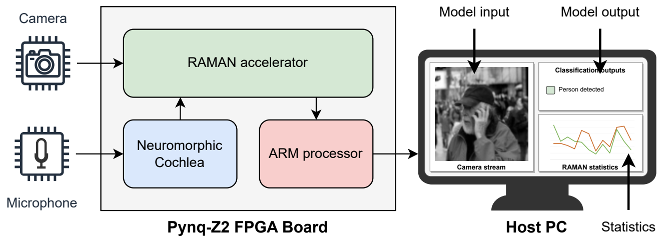

Fig. 1: The setup includes a host PC, camera and microphone sensors, and a Pynq-Z2 FPGA board. The neuromorphic cochlear model and RAMAN accelerator for neural network inference are deployed on the FPGA. The ARM processor on the FPGA sends the image received, cochleagram and the classified outputs to the PC to be visualized.

I. Demonstration Setup

We will demonstrate real-time audio and image classification using the RAMAN (Re-configurable and spArse tinyML Accelerator for infereNce) [1] TinyML accelerator. RAMAN is a TinyML accelerator that leverages quantization and sparsity in data and weights for efficient inference of convolutional neural networks on the edge.

For demonstrating image classification, an HM0360 sensor is used. It transmits grayscale images of 160x120 (QQVGA) resolution to the FPGA at 22fps using a MIPI CSI2 interface. The camera interface on the FPGA then selects a 96x96 window from the centre of each frame to be fed to the RAMAN accelerator.

For audio inference, a neuromorphic cochlea is used for efficient feature extraction. The neuromorphic cochlea incorporates a Cascade of Asymmetric Resonators (CAR) model [2], combined with a low-pass filter, to mimic the behavior of the Basilar membrane and Inner hair cells within the inner ear to extract frequency components in the audio efficiently. The audio is sampled by a microphone at 16kHz and is fed to the FPGA using an I2S interface [3]. The CAR model processes the audio from the microphone and produces cochleagrams which are fed to RAMAN for inference.

Depthwise separable convolutional neural networks operating on the RAMAN accelerator classify the images from the camera and the cochleagram from the cochlear model. For visual inference, a CNN trained on the Visual Wakewords (VWW) dataset is used which detects if a person is present in the frame. For audio inference, the model is trained on the google speech commands dataset for the keyword spotting (KWS) task classifies audio as one of 10 keywords. The ARM processor reads the inputs from the camera or cochlear model as well as the classified outputs using the AXI interface. The host PC uses an Ethernet interface with the ARM processor to receive the inputs and outputs of the network to be visualized.

II. Visitor Experience

We present an interactive experience where the visitors can view the live cochleagram from speech in the case of audio inference or the live video stream from the camera for visual inference, the model classification and various statistics of model inference.

III. Results

We will demonstrate the use of the RAMAN TinyML accelerator for real-time inference on both audio and visual modalities. The RAMAN accelerator, CAR model and the camera interface are deployed on a PynqZ2 board. The design runs at a maximum frequency of 50MHz and consumes 35k LUTs and 12.8k slice registers.

Real-life performance of neural networks deployed on the RAMAN accelerator for audio and visual modalities is tested demonstrating the feasibility of the RAMAN accelerator for edge machine learning inference.

Acknowledgements

The authors would like to express their sincere gratitude to their colleagues, Srikanth Rohit Nudurupati and Chandana D G for their invaluable help throughout this work.

References

[1] A. Krishna, S. R. Nudurupati, D. G. Chandana, P. Dwivedi, A. van Schaik, M. Mehendale, and C. S. Thakur, "RAMAN: A Re-configurable and Sparse tinyML Accelerator for Inference on Edge," *arXiv*, 2023. [Online]. Available: http://arxiv.org/abs/2306.06493

[2] Y. Xu, C. S. Thakur, R. K. Singh, T. J. Hamilton, R. M. Wang, and A. van Schaik, "A FPGA Implementation of the CAR-FAC Cochlear Model," *Frontiers in Neuroscience*, vol. 12, p. 198, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00198

[3] A. C. e. Adithya Krishna, Hitesh Pavan Oleti, "Live demonstration: Audio inference using neuromorphic cochlea on raman accelerator," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10389100

*Equal contribution.

This work is supported by the Pratiksha Trust under grant FG/SMCH-22-2106.