Collaborative Research: SWIFT: AI-based Sensing for Improved Resiliency via Spectral Adaptation with Lifelong Learning

Arjuna Madanayake (Florida International University FIU, 2229471), Sirani K. Perera (Embry Riddle Aeronautical University ERAU, 2229473), Francesco Restuccia (Northeastern University, 2229472), and Houbing Song (University of Maryland, Baltimore County UMBC, 2229473)

Introduction

This project is introduced to design and demonstrate a cutting-edge system for achieving spectral situational awareness through radio frequency (RF) machine learning (ML). The primary goal is to obtain actionable spectrum intelligence, a deep understanding of waveform characteristics, spectral content, and modulation techniques. Sub-6 GHz legacy band is considered as the main area of focus, which has gained significant attention due to recent FCC auctions around 3.5 GHz. This emphasizes the economic imperatives driving the need for robust access to legacy bands. The core objectives of the project will address an improvement factor of at least 10x real-time throughput over software-based signal awareness and spectrum sensing systems using new AI-chip technology and will enable cognitive radios with resilient, autonomous dynamic spectrum access.

Design goals and research objectives

Two themes : Theme A and Theme B.

- Theme A: Spectrum Adaptability: Network resilience through life-long learning at a contested and congested spectral interface.
- Theme B. Mixed-Signal Circuits and Components: Robust, energy-efficient, and high-performance AI chips for real-time lifelong learning at the radio-edge for real-time spectrum management and mitigation/measurement of harmful radio frequency interference (RFI) to passive

Wideband Multi-Beam Spectrum Sensor System Design

Possible approaches: design of AI spectral awareness sensor which addresses goals of Theme A and Theme B.

- Enhancement of bandwidth
- Reduce algorithm complexity
- Use of latest technologies

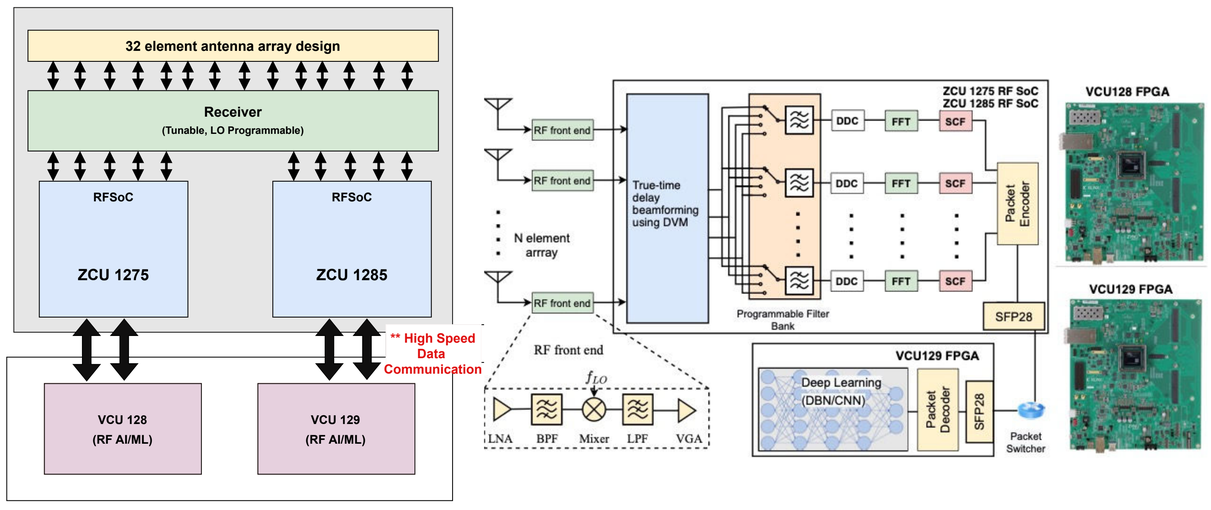


Figure 1: Proposed AI-accelerated Spectral Awareness Sensor

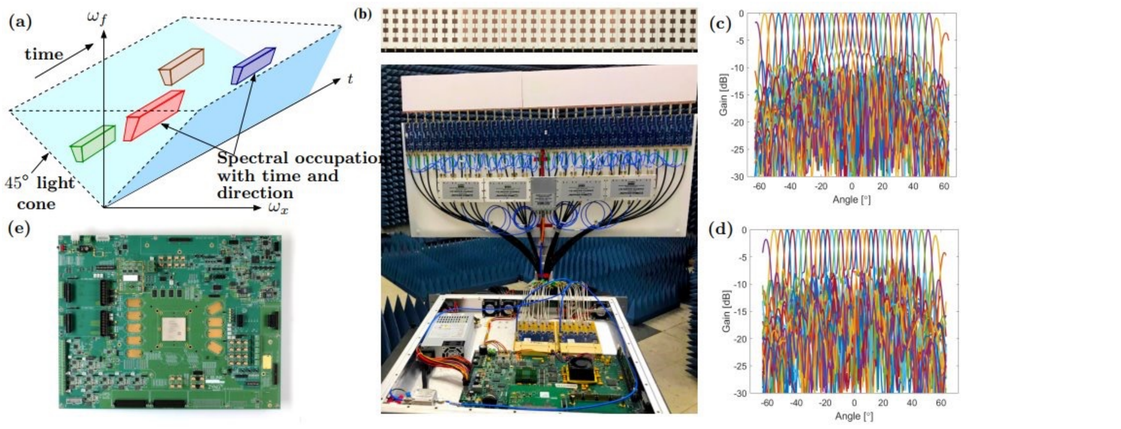


Figure 2: (a) Detection of spectral white spaces in 3-D (direction, time, frequency) space. (b) Experimen- tal setup of a 5.8 GHz 32-element approximate-DFT (ADFT) digital multi-beamformer, designed and built in Madanayake's RAND lab at FIU [1,2]. (c)-(d) Measured beam from (b): (c) ADFT, (d) fixed-point FFT. (e) Xilinx RF-SoC based ZCU-1275 platform at FIU.

[1] A. Madanayake, V. Ariyarathna, S. Madishetty, S. Pulipati, R. J. Cintra, D. Coelho, R. Oliviera, F. M. Bayer, L. Belostotski, S. Mandal, and T. S. Rappaport, "Towards a low-swap 1024-beam digital array: A 32-beam subsystem at 5.8 ghz," IEEE Transactions on Antennas and Propagation, vol. 68, no. 2, pp. 900–912, 2020.

[2] V. Ariyarathna, V. A. Coutinho, S. Pulipati, A. Madanayake, R. T. Wijesekara, C. U. S. Edussooriya‡, L. T. Bruton§, T. K. Gunaratne, and R. J. Cintra, "Real-time 2-d fir trapezoidal digital filters for 2.4 ghz aperture receiver applications," in 2018 Moratuwa Engineering Research Conference (MERCon), 2018, pp. 350–355.

Wideband Multi-Beam Spectrum Sensors with AI/ML Perception

Conventional methods use

- non-real-time processing
- high computational power

Current work: hardware architecture will enable real-time processing of incoming RF signals

Two level RF system on chip architecture to handle higher bandwidth

- Level 01 -RF level subbanding (16 GHz real-time within 24 GHz range)
- Level 02 – Fabric level parallelization

After level 01 subbanding, sampled signal with 1MHz bandwidth is made available across 8/16 channel

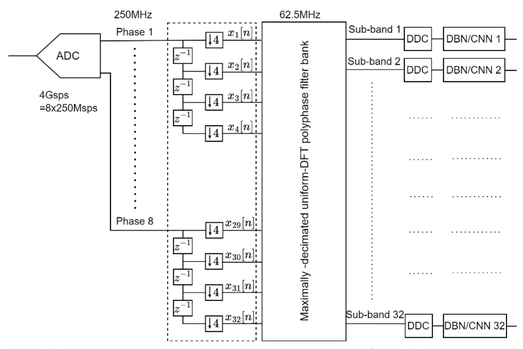


Figure 3: 2 level subbanding and maximally decimated uniform DFT polyphase filter bank

Proposed architecture -

- Use of maximally decimated uniform DFT polyphase filter bank - FIR filter bank with integrated fast Fourier transform for channelizing the 1GHz ADC signal to multiple sub bands
- Output of the filter bank - 32 channels
- AI ML algorithm is operating at 62.5 MHz clock

Deep belief network/Convolution neural network architectures will be used for spectrum intelligence - Extract information that helps improve spectrum utilization such as modulation type and direction of arrival.

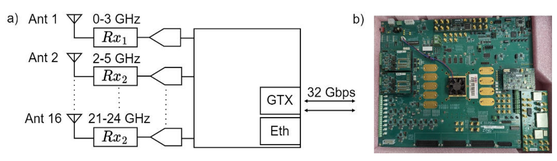


Figure 4: a) The 16 GHz band across FR1-FR3 consists of 16 1 GHz subbands chosen for AI/ML perception. The subbands are assigned to 16 optimized antennas (span 100 MHz-24 GHz); aggregate bandwidth is 16 GHz. Both ZCU-1275/1285 (16 channels) or ZCU-111 (8 channels) can be used for fast processing with 32 Gbps Serdes for connectivity to other systems. b) ZCU-1275 RF-SoC evaluation board

Implementing high speed data communication using multi-Gigabit transceiver links across multiple FPGAs

- Multi-Gigabit Transceivers/ Multi-Gigabit SERDES is a technology that receives parallel data and enables the transportation of high-bandwidth data over a serial link minimizing the number of I/O interconnects.
- Ability to connect multiple FPGA systems at high speed is useful for establishing multi-FPGA backends for transporting beamformed and processed channelized signal information.
- First stage completed: Established successful links with a line rate of 25 Gb/s connecting two Zynq UltraScale+ ZCU111 RFSoCs with the Virtex UltraScale+ VCU129 FPGA board using SFP28 connectors.

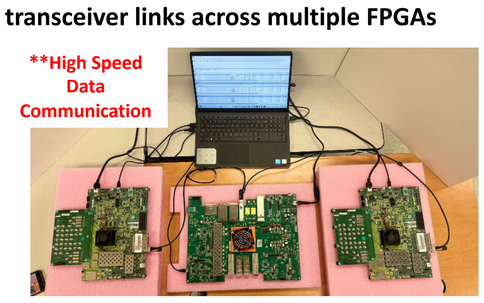


Figure 5: Experimental Setup: two ZCU111 RF SoCs connecting to a VCU129 board for implementing high speed data communication

Low-complexity Structured Neural Network (SNN) Architecture for Multi-beam Beamforming

A neural network architecture was proposed to realize multi-beam beamforming using structure-imposed weight matrices and submatrices.

- The structure as well as the sparsity of weight matrices and submatrices are shown to greatly reduce the space and complexity of the proposed network.
- The proposed neural architecture has $O(Mp^2L)$ complexity compared to a conventional fully connected L-layers of network with $O(M^2L)$ complexity, where M is the number of nodes in the input and output layers, p is the number of submatrices per layer, and M >> L, p.
- Numerical results show that the proposed architecture shows faster convergence without sacrificing the accuracy.

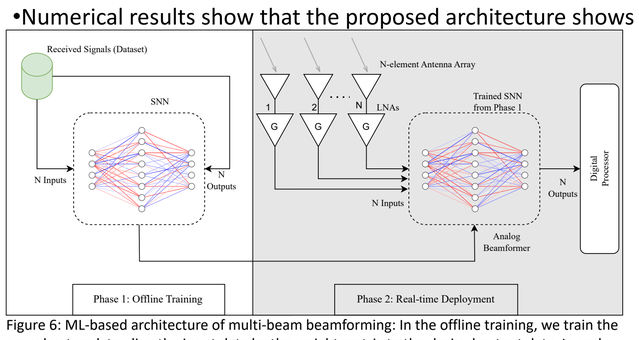


Figure 6: ML-based architecture of multi-beam beamforming: In the offline training, we train the neural network to align the input data by the weight matrix to the desired output data. In real-time deployment: RF signals from the antennas and low noise amplifiers (LNAs) are beamformed utilizing the structure imposed neural network, i.e., SNN. Once the multibeams are formed they will be sent to the digital processor

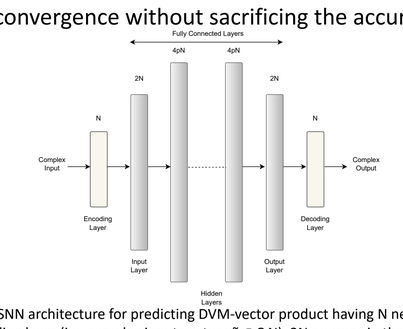


Figure 7: SNN architecture for predicting DVM-vector product having N neurons in the encoding layer (i.e. complex input vector $x^{\sim} \in C\ N$), 2N neurons in the input layer (separating real-vales and imaginary parts of the input vector $x^{\sim}$ giving 2N neurons in the input vector $x \in R\ 2N$), 4pN neurons in the hidden layer, where p, $N \in Z$ + and p 2 submatrices appearing in the weight matrices between hidden layers, and 2N neurons in the output vector $y \in R\ 2N$ resulting the beamformed vector $y^{\sim} \in C\ N$.

Forward & Back propagations of the SNN for Multibeam Beamforming

Proposition II.1 An L layers supervised structured neural network (SNN) can be introduced via weight matrices $\{W^{(1)}, W^{(1)}, \cdots, W^{(l+1)}\}$ among l number of hidden layers, where $l = L-2$. The weight matrices are defined via $W^{(1)} = \begin{bmatrix} w_1^{(1)} & \cdots & w_p^{(1)} \end{bmatrix}^T$ having submatrices $w_i^{(1)}$ for $i = 1, 2, \cdots, p$ calculated as a product of sparse matrices s.t. $\check{D}_{2M} F_{2M} J_{2M\times 2N} \hat{D}_{2N}$; $W^{(l)} = [w_{i,k}^{(l)}]_{i,k=1}^{p}$ with submatrices $w_{i,k}^{(l)} = I_{2M}$; and $W^{(l+1)} = [w_1^{(l+1)}, w_2^{(l+1)}, w_3^{(l+1)}, ..., w_p^{(l+1)}]$ with submatrices $w_i^{(l+1)}$ calculated as a product of sparse matrices s.t. $\hat{D}_{2N}[J_{2M\times 2N}]^T F_{2M}^*$.

We note here that $\alpha = e^{-j\omega\tau} \in \mathbb{C}$, $N = 2^r (r \geq 1)$, ω is temporal frequency, and τ is a time delay, $M = 2N$, $\hat{D}_N = diag[\alpha^{\frac{k^2}{2}}]_{k=0}^{N-1}$, $J_{M\times N} = \begin{bmatrix} I_N | & 0_N \end{bmatrix}^T$, I_N is the identity matrix, 0_N is the zero matrix, $F_N = \frac{1}{\sqrt{N}}[\omega_N^{kl}]_{k,l=0}^{N-1}$ is the DFT matrix with nodes $\omega_N = e^{-\frac{2\pi j}{N}}$, $j^2 = -1$, and F_M^* is a conjugate transpose of F_M.

Proposition II.3 Let the objective function of the proposed SNN be given by $O(W^{(1)}, \cdots, W^{(l+1)}) = \frac{1}{NM_b} \sum_{s=1}^{M_b} \sum_{k=1}^{N} (y_s^{(k)} - \hat{y}_s^{(k)})^2$, where $W^{(1)}, W^{(2)}, \cdots, W^{(l+1)}$ are the weight matrices defined among l hidden layers of the network. Then, the loss function O with respect to W^{l+1} is given via $\frac{\partial O}{\partial y^l} = \left(\frac{\partial O}{\partial y^{l+1}}\right) \cdot \left(\frac{\partial y^{l+1}}{\partial a^{l+1}}\right) . W^{l+1}$ and $\frac{\partial a^{l+1}}{\partial w_i^{l+1}} = y_i^l$, where $a^{l+1} = W^{l+1} . y^l + \theta^{l+1}$, $y^{l+1} = \sigma(a^{l+1})$, $\sigma(.)$ is the activation function of the current layer, and W^{l+1} is determined by submatrices w_i^{l+1}.

Numerical Results of the SNN model

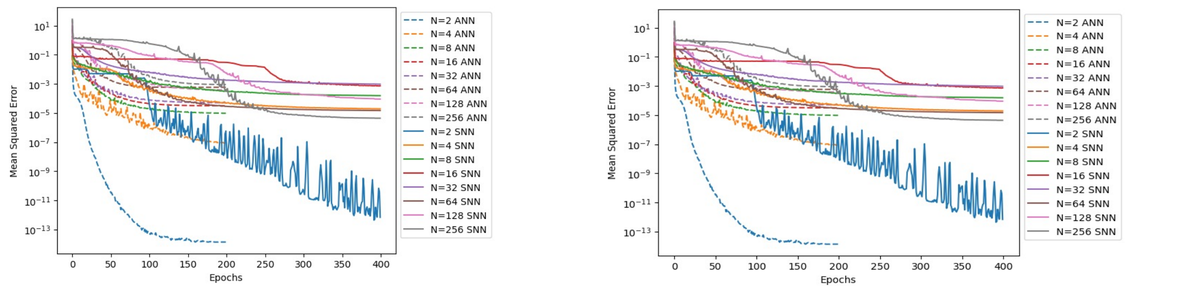
N	Model/ Weights(ANN)	MSE (ANN)	Model/ p/ λ/ Weights(SNN)	MSE (SNN)	Pr(Weights)
2	$(4, 8, 4)/76$	7.3790×10^{-14}	$(4, 8, 4)/1/2/76$	1.1183×10^{-10}	0%
4	$(8, 16, 8)/280$	9.8138×10^{-11}	$(8, 16, 8)/1/3/168$	2.5357×10^{-5}	40%
8	$(16, 32, 16)/1072$	9.6169×10^{-6}	$(16, 64, 16)/2/4/344$	3.9557×10^{-4}	67%
16	$(32, 64, 32)/4192$	1.8101×10^{-4}	$(32, 64, 32)/1/2/1344$	7.1806×10^{-4}	68%
32	$(64, 128, 64)/16576$	2.9902×10^{-4}	$(64, 128, 64)/1/2/4736$	1.0074×10^{-3}	71%
64	$(128, 512, 128)/131712$	6.7401×10^{-4}	$(128, 512, 128)/2/2/35200$	7.7334×10^{-5}	73%
128	$(256, 1024, 256)/525568$	4.8125×10^{-4}	$(256, 1024, 256)/2/2/135936$	4.3494×10^{-4}	74%
256	$(512, 4096, 512)/4198912$	1.2377×10^{-3}	$(512, 4096, 512)/4/2/1067520$	6.5466×10^{-4}	75%

TABLE I: This table shows MSE values having different elements of antenna arrays. These values are obtained using codes written in Python (Version-3.10) along with the TensorFlow (version-2.0) framework, and compiled with Adam optimizer. The term "Model" consists of three numbers representing nodes in input, hidden, and output layers. The notations p and λ denote the number of submatrices and recursive steps, respectively. The last column shows the percentage of savings on utilizing SNN over ANN.

N	FLOPs(ANN) (Simulation)	FLOPs(SNN) (Simulation)	FLOPs(SNN) (Eq.(15) + Eq.(16)	Pr(FLOPs)
2	136	132	132	3%
4	528	312	312	55%
8	2080	1440	1440	46%
16	8256	4640	4640	44%
32	32896	17472	17472	47%
64	262656	135424	135424	48%
128	1049600	532992	532992	49%
256	8392704	4229120	4229120	50%

TABLE II: Addition and Multiplication counts(FLOPs) for the SNN and fully connected neural network, i.e., FLOPs: = #a(SNN) + #m(SNN). The second and third column values are obtained using codes written in Python (Version3.10) along with the TensorFlow (version - 2.0) framework. We note here that the third and fourth columns show the same numerical values, and hence conclude the coincident of the theoretical results in equations (15) and (16) with the numerical simulations. The last column shows the percentage of the savings on utilizing SNN (executing λ < r recursive steps) over ANN.

Performance and Validation Results of the SNN



a) Training Performance

b) Validation Performance

Figure 8: The figures (a) and (b) show training and validation results of the SNNs based on the different elements of antenna arrays, i.e., N = 2, 4, 8, · · · , 256. These graphs are obtained referencing the "Models" listed in Table I. When training ANN and SNN models for 200 epochs, they converge to MSE values of 10−4 and 10−2, respectively. This shows that there is a challenge in maintaining the complexity and accuracy simultaneously. Thus, to obtain the MSE with the accuracy of 10−4, we trained the SNN for 400 epochs. These graphs are obtained using *Python (Version-3.10)* along with the *TensorFlow (version- 2.0)* framework and compiled with *Adam* optimizer.

AI-Driven Semantic Spectrum Segmentation

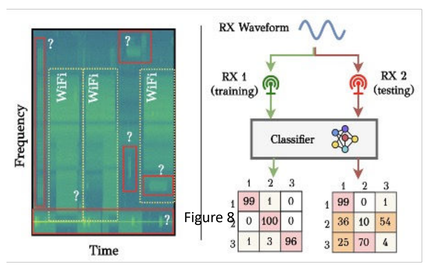


Figure 8

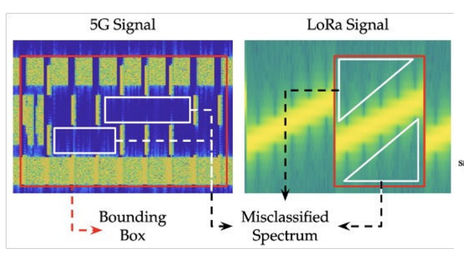


Figure 9

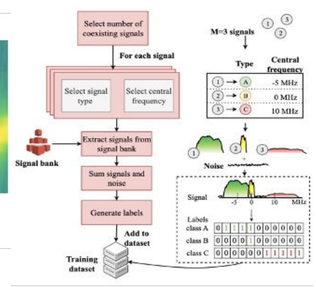


Figure 10

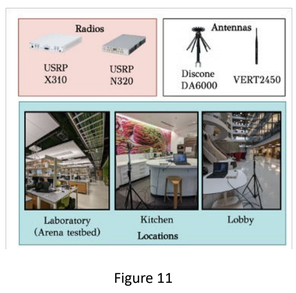


Figure 11

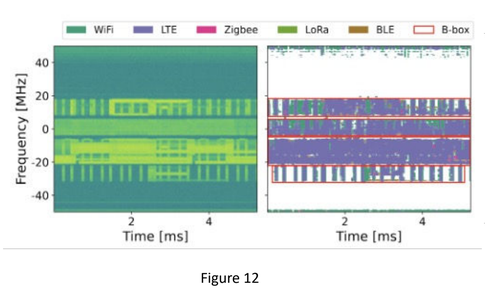


Figure 12

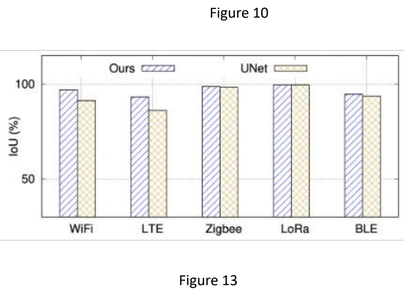


Figure 13

Reference: D. Uvaydov, M. Zhang, C. Robinson, S. D'Oro, T. Melodia and F. Restuccia. "Stitching the Spectrum: Semantic Spectrum Segmentation with Wideband Signal." IEEE INFOCOM 2024.

Publications

- S. M. Perera, G. Rathnasekara, and A. Madanayake, "Thiran Filters for Wideband DSP-Based Multi-Beam True Time Delay RF Sensing Applications," Sensors 2024, MDPI, vol. 24(2), 2024
- H. Weerasooriya, G. Rathnasekara, F. Restuccia, and A. Madanayake, "RF-SoC Platforms for RF-AI Spectrum Perception," in 2024 International Applied Computational Electromagnetics Society (ACES) Symposium, IEEE, 2024 (paper accepted).
- K. Karunanayke, H. Weerasooriya, G. Rathnasekara, A. Singh, T. S. Rappaport, J. M. Jornet, and A. Madanayake, "Design of 145 GHz BPSK Modem on RF-SoC", in 2024 International Applied Computational Electromagnetics Society (ACES) Symposium, IEEE, 2024 (paper accepted).

Broader Impacts

- PI Restuccia is expanding the Young Scholars Program for K-12 students and UPLIFT program for undergraduate researchers at Northeastern university
- 1 Journal paper and 2 conference paper publications
- 2 female PhD students are working on the research