Vimal William

https://vimalwill.github.io/

Email : vimalwilliam99@gmail.com
Mobile : +91-638-194-4471

RESEARCH INTERESTS

HW/SW co-design, Domain-specific accelerators, Compiler optimization, Hardware-aware AI Optimization, Deep Learning model compression

EDUCATION

Anna University — Chennai, India
Bachelor of Engineering in Electronics and Communication; (CGPA: 8.5/10) — *Aug. 2019 – Mar. 2023*

RESEARCH EXPERIENCE

- **Low-Power Hardware Systems Group** — Bangalore, India
  Researcher (Advisors: Dr. Arnab Raha (Intel US), Dr. Debayan Das(IISc India)) — *Jan 2023 - Present*
  - **NLAE**: Actively contributing to the Algorithm/Hardware co-design for the Non-Linear Activation Approximation Engine which is part of a domain-specific accelerator for the SOTA AI model dedicated to brain cancer detection. Working in the Taylor-series-based approximation for activation functions employed in CNNs and Transformers and developed an algorithm that calculates the length of the Taylor-series based on the sensitivity of the activation function. Additionally, involved in the research for NN compression and Hardware aware-AI optimization.

PUBLICATIONS

1. Vimal W and Akshansh Gupta. Deep learning architecture for motor imaged words. *arXiv e-prints*, pages arXiv–2308, 2023
2. Vimal W. Study on the behaviour of mel frequency cepstral coefficient algorithm for different windows. In *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6, 2022

WORK EXPERIENCE

- **SandLogic Technologies** — Bangalore, India
  System Software Engineer (Edge AI & Hardware) — *Mar. 2023 - Present*
  - **SL-DLA**: Working on an MLIR-based back-end and codegen for a re-configurable Deep Learning Accelerator (DLA) and actively contributing to the optimizations to reduce the computation overheads in the accelerator. Helps Hardware engineers in mapping AI operators and handling precision such as Int8, Int4 and Fp8 in accelerator design
  - **SL-NN Optimizer Toolkit**: Developed post-training structured pruning algorithms with various salience estimation methods for effectively compressing Neural Networks for edge deployment. The algorithm supports numerous AI layers such as the Attention, linear, and Convolution layers. The compressed Neural Networks exhibit a better inference rate with minimal power consumption. Actively contributed to developing SL Deep Compression (pruning & quantization) library and worked on INT8 and INT4 quantization algorithms. Working on research to reduce the outliers due to the effect of quantization in LLMs.
  - **EEMMCC**: Designed and developed the Embedded Electronics Module for the Multiscale Computational Camera, specializing in the long-distance streaming pipeline. Utilized GStreamer and harnessed Zynq Ultrascale's VCU for high-performance hardware-accelerated video encoding.
  - **GstPlugin**: Trained engineers in constructing a GStreamer plugin for hardware-aware acceleration for image processing algorithms and achieved an execution rate of 1ms in CPUs and 0.4ms in GPUs
- **SandLogic Technologies** — Bangalore, India
  Edge AI Intern — *Sep. 2022 - Mar. 2023*
  - **SL-DLA**: Developed packed-SIMD-based Deep Learning operators for RISC-V CPU which enables the runtime to accelerate the non-supported operator fallback from DLA to CPU and improved the inference rate

- **Robert Bosch Center for Data Science & Artificial Intelligence** Chennai, India
  Research Intern *June. 2022 - Aug. 2022*
  - **Deployable AI**: Contributed to the project titled "Deployable AI for Smart Buildings: Transfer Learning Method Development for DL-Based Controls in HVAC" under the mentorship of Dr. Satyanarayanan Seshadri (Professor, IIT Madras).
- **Central Electronics Engineering Research Institute, CEERI - CSIR** Pilani, India
  Research Intern *Jan. 2022 - June. 2022*
  - **BCI**: Developed a hybrid Deep Learning model with the combination of convolution layers (Conv2D & Conv1D) with the LSTM layer to enhance the translation of neural signals from the motor cortex region. The AI system includes SOTA signal processing techniques which include wavelet-based de-noising, matrix decomposition, and Gramian Angular Field (GAF) for noise reduction in neural signals. The AI model performs a signal translation with 97% accuracy
- **National Institute of Ocean Technology** Chennai, India
  Research Intern *Jan. 2021 - March. 2021*
  - **Low-Power Inference**: Worked in the Development of a Deep Learning model for amplitude estimation of waves and compressed the model for low power inference at micro-controllers under the mentorship of Mr. Nitesh Varma (Project Scientist, NIOT) and Dr. Bolem (Scientist - E, NIOT).

PERSONAL PROJECTS

Iris Engine GitHub

Deep Learning Accelerator for Neural Radiance Field (NeRF) developed for medical imaging. The accelerator was designed and mapped to FPGA with Vitis HLS libraries and interfaced with PS through XRT (Xilinx Runtime) C++ APIs.

VStream - Video Analytics Pipeline GitHub

GStreamer-based streaming pipeline enabled with YOLOV5-based object detection model. The pipeline offloads the workload to either the CPU or GPU through OpenCV libraries based on the inputs from the OpenCL APIs.

HONOURS & AWARDS

1. Contributory Talk on "Deep Learning Architecture of Motor Imaged Words" at "5th International Conference on Recent Advances in Mathematical Science with Applications in Engineering and Technology".
2. 7th President of RAC Jerusalem, Chennai (2022-23).
3. Change Maker's Award from District Rotaract Council - 3232 (2023)

SKILLS

Frameworks: TensorFlow, PyTorch, ONNX OpenVino, XLA/IREE, MLIR, LLVM, OpenCL, GStreamer, Vitis

Programming: C & C++, Python, Bash