Saikiran Chepa | DSP Engineer | C & Assembly | Signal Processing | AI & ML

Hyderabad, India

+91 8977890154 • sai.406@gmail.com

saikiran-chepa-81938449/ • chbsaikiran

Professional Summary

Results-driven AI/ML Engineer with 5+ years of hands-on experience in deep learning and large language model development, complemented by a strong background in embedded DSP optimization. Proficient in training and deploying models including CNNs, RNNs, transformers, ResNet, and LLMs, with practical expertise in pre-training, supervised fine-tuning (SFT), and RLHF. Prior 11+ years of expertise in optimizing Speech and Audio signal processing algorithms across ARM, QDSP, Intel, RISC-V, SHARC, and TI platforms offers a unique blend of AI innovation with high-performance embedded computing.

Key Highlights

- Expert in Digital Signal Processing with deep knowledge of FIR/IIR filters, multirate signal processing, and matrix multiplication optimizations.
- Optimized DSP algorithms for multiple architectures, including RISCV, SHARC, Intel AVX2, ARMV8-A, and ARM NEON intrinsics, significantly reducing execution time.
- Proficient in floating-point to fixed-point conversions for DSP applications, improving computational efficiency across different hardware platforms.
- Developed scripts for automated performance profiling using Python, Simpleperf, and ADB commands to analyze MCPS performance on Android and Raspberry Pi 4.
- Implemented FFT and IFFT optimizations using RISCV and SHARC DSP libraries.
- Hands-on experience in AI/ML model training on large datasets, including ResNet-50, smolLM2, and LLMs, with model training and deployment on AWS and Hugging Face.

Education

- **MTech in Communication Systems**, IIT Madras
- **BTech in Electronics and Communication Engineering**, GRIET, JNTU Hyderabad

Experience

Engineer III, Capgemini, Client Qualcomm Hyderabad

Dec 2016 – May 2023 | Jan 2025 – Mar 2025

- Floating point to Fixed Point conversion. Implementing log10 and pow using fixed point basicops functions log2 and pow2.
- Using Matlab's codegen to convert Matlab code to C code.(Making changes to original Matlab code so as resolve codegen errors). Dynamically increasing arrays in matlab code were implemented using fixed size arrays.
- Good understanding of writing CSIM APIs for a Matlab code.
- Coded FIR filter followed by decimation by 2 in assembly, by not calculating the alternate samples. The Left and right channel of input used the same FIR coefficients so in the inner most loop of FIR filter both left and right outputs were calculated together this allowed all slot in the packets to be utilized. Also implemented up sampling followed by FIR filter.
- Batch files, Shell scripts, python scripts and cmake to automate compiling and testing of code.
- Good Understanding about Digital Signal Processing theory. FIR and IIR filters, Multirate signal processing.
- Optimized Matrix Multiplication function for AGVC(Advanced Generative Voice Coder) for Intel AVX2 architecture. AGVC uses AI/ML for speech compression.

DSP Engineer , Capgemini, Client Goodix Bangalore

Aug 2023 – Jul 2024

- Used RISCV DSP Lib's fft and ifft function and integrated them into VoiceExperience Code. The code will not be BitExact, so validated the outputs using matlab scripts which generated rmse in dB and also checked the output wav files ADOBE Audition.
- SHARC is floating point DSP, the fft and ifft library functions were available only in floating point. Integrated these float codes into CarVoice code by converting fixed point input to float and using fft and ifft lib function and then again converting there outputs back to fixed point.
- Optimized loops of function by converting input to float and doing processing in float and then converting back to fixed point. This helped in reducing cycles by 60 percent, compared to just using fixed point code.
- Optimized CarVoice Code for ARMV8A architecture, brought the factor between ARMV7A optimized code and ARMV8A optimized from factor of 2.2 to 1.4. For this coded 25 functions in arm neon intrinsics. Wrote python scripts to compare ARMV7A flat profile with ARMV8A flat profile and arrive at functions not optimized for ARMV8A.
- Obtained MCPS on android device using adb commands, used simpleperf for getting flat profile on android device.
- Worked on Raspberry Pi 4, and generated MCPS for ARMV8A and ARMV7A code. For this wrote scripts in python that would extract minimum MCPS out of 5 runs of the executable for each frame and then get average and peak of these minimums.

Senior DSP Engineer , Couth Infotech

Apr 2012 – Mar 2015

- Optimisation on C66x & C64xplus of SILK, Speex, Opus Codecs, Linear assembly code was written.
- Scratch and Channel were split into blocks of 4KB.
- Stack usage was reduced by using scratch based implementation.
- NDK setup was done for C6678 and C6472 boards to test the codec for standard & Non-Standard test vectors.
- Wrappers written for testing are
  - DataMove(Channel & Tables run time relocatable).

- Illegal read/write.
- Register preservation, Input buffer corruption, Input/Output buffer alignment.
- Interrupt testing.
- Stack and buffer calculation.
- Code Coverage.
- Scratch contamination

Senior Software Engineer , Aricent

Jul 2005 – Jul 2009

- eAAC plus Decoder Optimizations on C64xplus
- Interleaved and non-interleaved output support was developed.

Certifications

- Domain Scholar Certificate in Signal Processing and Communications (NPTEL, 60-week program)
  - Courses Done : Discrete Time Signal Processing, Applied Linear Algebra, Probability Foundations for Electrical Engineers, Multirate DSP, Introduction to Information Theory, Principles of Signals and Systems, Adaptive Signal Processing
- Deep Learning Specialization – Coursera
  - Gained expertise in designing, training, and optimizing deep neural networks, including CNNs, RNNs, and transformer-based models using TensorFlow.
  - Applied advanced techniques such as neural style transfer, object detection, and recognition for image/video tasks.
  - Implemented vectorized deep learning architectures, optimized hyperparameters, performed bias-variance analysis.
- ERA V3 (Extensive Reimagined AI, Version 3) – The School of AI, Bangalore
  - Trained full-scale ResNet models on the complete ImageNet dataset from scratch, and developed small-scale Large Language Models (LLMs) with hands-on experience in pre-training, supervised fine-tuning (SFT), and Reinforcement Learning with Human

Feedback (RLHF).

- Capstone Project: Built an AI agent capable of answering natural language queries about Gmail account data (e.g., annual spending on Zomato, travel bookings via Redbus), demonstrating practical integration of Google Gemini and AI agents for personalized information retrieval.

Technical Skills

- **Programming Languages:** C, Python, MATLAB
- **Platforms:** ARM, TI, Intel, QDSP, SHARC, RISCV
- **IDEs:** Visual Studio, CCS, Eclipse, DS-5, Spyder, ARM Workbench, Cursor AI, VS Code, AWS
- **Environment:** Windows, Linux
- **Version Control:** GIT, SVN