Ramakrishnan Sivakumar

Senior Machine Learning Software Engineer, Seattle WA

An ML software engineer with 7+ years of experience, designed and developed deep learning frameworks and operating systems with expertise in system architecture, power and performance optimization, and close collaboration with customers.

✉ ramkrishna2910@gmail.com

ramkrishna2910.github.io/portfolio

linkedin.com/in/ramakrishnansivakumar

medium.com/@ramkrishna2910

WORK EXPERIENCE

Senior Machine Learning Software Engineer
Groq Inc

09/2021 - Present

Design and development of Groq's Tensor Streaming Processor based solutions for real world machine learning workloads. Co-engineer the Groq software stack with learnings from real world deployment of Groq hardware and software focused on improving developer velocity and experience.

Deep Learning Software Engineer
Intel Corporation

08/2018 - 09/2021

Development and optimization of Onnxruntime and WinML frameworks to enhance end-to-end user experience on mobile systems.

Operating System Software Engineer
Intel Corporation

07/2016 - 08/2018

Development of Windows Operating system to improve power and performance of Intel platforms.

KEY ACCOMPLISHMENTS

Graph Neural networks on Groq Architecture

- Identified GNNs as a good fit for the Groq architecture and enabled their implementation on this platform. Designed Scatter/ Gather algorithm for the deterministic Groq architecture and collaborated with Argonne National Labs and Oak Ridge National Labs on graph classification problems for drug discovery and material science. Achieved up to 40x better performance compared to state-of-the-art GPUs.

GroqFlow

- Developed a high-performance inference pipeline for the Groq ecosystem. Designed and implemented the framework's model processing pipelines, execution model, and integration with other frameworks and libraries. GroqFlow is an efficient and scalable software solutions for a novel AI computing system that is widely used by customers for interacting with Groq products.

MLAgility

- Developed an end-to-end machine learning platform that streamlined building and benchmarking models on CPUs, GPUs and Groq architecture. Designed and implemented the platform's data pipelines, model preprocessing workflows, and deployment infrastructure. MLAgility has been open sourced and democratizes access to broad and reliable benchmarking information across AI hardware ecosystems.

Development of ONNX and ONNXRuntime

- Designed and developed low level matrix multiplication kernels for acceleration of Int8 inference on CPUs. Implemented platform aware matrix blocking to optimally use the cache hierarchy and custom thread management for hybrid CPUs to maximize throughput and minimize single inference latencies. Enabled frameworks to leverage hybrid architecture of CPUs to maximize throughput across multiple cores with varying performance capabilities
- Active contributor to the ONNX specification

PUBLICATIONS

Patent

Technology to augment thread scheduling with temporal characteristics

02/09/2021

Blog

Optimizing BERT model for Intel CPU Cores using ONNX runtime default execution provider

03/01/2021

Academic Paper

A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads

09/13/2022

TECHNICAL SKILLS

Programming Languages

C, C++, Python, X86 assembly

Frameworks

OnnxRuntime, OpenVino, PyTorch

EDUCATION

Master's in Computer Engineering
University of North Carolina at Charlotte, NC

08/2014 - 05/2016

4.0