Lakshmi Kavya Kalyanam

GPU/ML Engineer Specializing in Neural Networks
Tampa, FL · +1-813-609-9796 · kavyakalyanam@gmail.com
LinkedIn · GitHub

Professional Summary

Machine Learning Engineer with 5+ years experience in neural network compression, CUDA optimization, and edge computing. Author of 9 peer-reviewed publications and 3 patents in model compression and GPU optimization. Passionate about advancing scalable, hardware-efficient AI solutions.

Skills

Programming Languages: Python, C++, CUDA, Embedded C
ML/AI Frameworks: PyTorch, TensorFlow, ONNX, scikit-learn
Optimization Techniques: Structured/Unstructured Pruning, Quantization, Dynamic Sparsity (RigL)
Hardware Platforms: Arduino, FPGA
Tools: Docker, Git, Linux

Professional Experience

Research Engineer and PhD Fellow — University of South Florida

Tampa, FL Jan 2019 - Present

- Architected 5-stage MLLR-inspired compression pipeline using PyTorch 2.0+cu124 on NVIDIA GTX 1080 Ti
- Designed activation-aware FP16 model compression achieving 82• Built distributed object detection framework using Darknet and YOLOv4 on ARM Cortex-M4 microcontrollers
- Developed ONNX graph rewriting algorithms for sparse model deployment, enabling 60

Research & Publications

S. Boyidapu, L. K. Kalyanam and S. Katkoori (2024). "Automated Hidden Neuron Optimization for Multilayer Perceptrons for Classification Tasks." *International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV).*
L. K. Kalyanam, S. Katkoori (2023). "Unstructured Pruning for Multi-Layer Perceptrons with Tanh Activation." *IEEE International Symposium on Smart Electronic Systems (iSES)*, DOI: 10.1109/iSES58672.2023.00025.
L. K. Kalyanam, S. Katkoori (2023). "Sigmod-based Neuron Pruning Technique for MLPs on IoT Edge Devices." *International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV)*, DOI: 10.1109/ICPEEV58650.2023.10391875
Kalyanam, L.K., Joshi, R., Katkoori, S. (2023). "Layer-Wise Filter Thresholding Based CNN Pruning for Efficient IoT Edge Implementations." *IFIP Advances in Information and Communication Technology*, Springer, Cham.

Key Projects

Distributed Real-Time Object Detection Framework

Methodology: Advanced ML optimization — **Tools:** PYNQ Z1 AP-SoC (Xilinx Zynq), Binarized Neural Networks (BNNs), Embedded Computer Vision
Built a scalable, low-latency distributed system using PYNQ-Z1 AP-SoCs for real-time object detection via Binarized Neural Networks (BNNs), achieving 19.23 FPS across 3 edge nodes.

- Heterogeneous computing on FPGA-based SoCs

Dynamic Sparsity Optimization for CNNs

Methodology: Compiler optimization — **Tools:** PyTorch 2.6.0+cu124, CUDA 12.3, NVIDIA GTX 1080 Ti
Designed a modular, compiler-aware optimization pipeline for compressing and deploying convolutional neural networks (CNNs), achieving over 98• Achieved 98.98

Neural Network Optimization Framework

Methodology: Neural network pruning — **Tools:** Deep Neural Networks (DNNs), ONNX Runtime and Graph Transformations, Arduino and Embedded Microcontrollers
Engineered a unified, correlation-based pruning framework for deep neural networks (DNNs) applied across 9 diverse datasets, incorporating activation-aware optimization, ONNX export, and real microcontroller deployment for end-to-end benchmarking and deployment guidance.

- Reduced model size by 75

LitBot: AI Literature Survey Assistant

Methodology: Advanced ML optimization — **Tools:** OpenAI GPT models, FAISS vector search, PyMuPDF PDF processing
Developed LitBot, a GPT + FAISS-powered AI assistant for automating literature surveys, semantic search, and citation anal-

ysis across 200+ research papers, streamlining academic workflows.

- Successfully integrated semantic chunking, summarization, and citation linking a...
- AI/ML-specific document preprocessing and semantic chunking

Resume Editor Bot

Methodology: Advanced ML optimization — **Tools:** Python 3.9+, FastAPI, LangChain
An intelligent resume editing assistant powered by Retrieval-Augmented Generation (RAG) and large language models (LLMs), designed to help users create, optimize, and tailor their resumes dynamically based on job descriptions. Features project-based resume generation, smart project ranking, and job-specific content tailoring.

Education

Ph.D. in Computer Science — *University of South Florida* Expected 2025
Focus: Sparse model optimization, embedded ML, neural network compression
Achievements: 9 peer-reviewed publications, 3 patents filed, 2 Best Paper Awards

Master of Science in Computer Science — *University of South Florida* 2021
Focus: Machine Learning, Computer Vision, Embedded Systems
GPA: 3.9/4.0