Luis Daniel Ferreto Chavarría

☎ +506-60418618 • TheRadDani • LinkedIn • ldanielfch@gmail.com • Portfolio
Alajuela, Costa Rica

Machine Learning Researcher with a focus on deep learning, reinforcement learning, and computer vision. Passionate about developing efficient AI models for edge computing and hardware acceleration, seeking research opportunities to advance the state-of-the-art in AI.

Education

Costa Rica Institute of Technology — Cartago, Costa Rica
M.S.c., Computer Science — *May 2023 – Present*

University of Costa Rica — San José, Costa Rica
B.S.c., Electrical Engineering — *Mar 2018 – Dec 2022*

Work Experience

Intel — Oct 2024 – Present
AI Frameworks Engineer — *Heredia, Costa Rica*

- Developed and optimized deep learning applications using frameworks such as PyTorch, TensorFlow, and ONNX Runtime, focusing on improving performance and efficiency for modern CPU/GPU architectures.
- Conducted performance analysis, profiling, and optimization for deep learning models, achieving substantial gains in latency and throughput by implementing techniques like model compression, quantization, and graph compiler optimizations.
- Utilized CUDA programming to accelerate key deep learning operations and kernels, optimizing for hardware architectures including Intel Xeon processors and GPUs.
- Integrated Neural Architecture Search (NAS) techniques to enhance model performance, exploring hardware-aware optimizations for deep learning frameworks.

Hewlett Packard Enterprise — Aug 2023 – Oct 2024
Systems Software Engineer — *Heredia, Costa Rica*

- Specialized in low-level hardware and embedded programming using C, Yocto, and Linux Kernel development, focusing on advanced storage protocols (SSDs, HDDs, NVMe, SCSI, PCIe).
- Automated system operations using Python and Bash scripting for Linux environments.
- Implemented scalable virtualization solutions using Docker containers and Kubernetes.
- Developed containerized applications for API interactions with data servers, utilizing Rust for efficient programming and system performance.

Costa Rica Institute of Technology — Apr 2023 – Present
Postgraduate Researcher — *Cartago, Costa Rica*

- Collaborated with undergraduate and postgraduate researchers to enhance deep learning model performance on FPGAs through hardware acceleration, model compression, quantization, and distillation.

Walmart Global Tech Jul 2022 – Jul 2023
Analytics Engineer *Heredia, Costa Rica*

- Optimized operational efficiency by analyzing and visualizing supply chain data, conducted A/B testing and exploratory data analysis (EDA).
- Developed and implemented scalable machine learning models (Gradient Boost, Bayesian Optimization, Isolation Forests) and data processing to enhance predictive analytics and decision-making.

Research Interests

My research centers on optimizing deep learning models for resource-constrained environments through techniques like quantization and pruning. I am exploring **reinforcement learning**, particularly **TinyRL**, for complex tasks such as multi-object tracking. Additionally, I am investigating **hardware-aware neural networks** to enhance computational and energy efficiency in edge computing. Furthermore, I am exploring the potential of **neuromorphic computing** and **generative AI** to revolutionize hardware design and accelerate AI advancements.

TinyRL aims to achieve real-time, robust multi-object tracking on resource-constrained devices, surpassing traditional computer vision methods in adaptability and efficiency.

Hardware-Aware Neural Networks develops neural networks seamlessly integrated with hardware, maximizing computational efficiency and energy savings.

Neuromorphic Computing explores brain-inspired architectures to achieve unprecedented energy efficiency and real-time performance in AI applications.

Generative AI accelerates hardware design and optimization through AI-driven automation, leading to more efficient and innovative systems. Using quantization and pruning to optimize transformer models for deployment on edge devices by reducing computational costs without sacrificing performance.

Certifications

- Generative AI with Large Language Models by AWS and DeepLearning.ai
- Production Machine Learning Systems by Google
- Neural Networks and Deep Learning by DeepLearning.ai
- DeepLearning.AI TensorFlow Developer by DeepLearning.ai
- Introduction to Concurrent Programming with GPUs by Johns Hopkins University

Skills

- **Data Engineering Tools:** Apache Airflow, Hadoop, Spark.
- **Cloud Platforms & Deployment:** AWS (EC2, S3, SageMaker), Docker, Google Cloud Platform, Kubernetes.
- **ML Frameworks:** Keras, Neo4j, NLTK, OpenCV, PyTorch, Scikit-Learn, TensorFlow.
- **Programming Languages:** C/C++, CUDA, Python, R.
- **Version Control & Collaboration:** Git, GitHub, GitLab.

Languages

- **Spanish** (Native)
- **English** (C1)
- **German** (A1b)
- **Italian** (Intermediate)