Hamza Elshafie

Linkedin: https://linkedin.com/in/hamzaelshafie
Github: https://github.com/HamzaElshafie
Email: hamzaelshafie@outlook.com
Work Blog: hamzaelshafie.bearblog.dev

UK Mobile: +447765949651
EG Mobile: +201001660400
Address: Cairo, Egypt — Open to Relocation

Professional Summary

Machine Learning Engineer with a Master's in Machine Learning and a Computer Science background. Skilled in deep learning, with experience developing and optimising AI models using PyTorch and 3+ years of Python experience. Focused on low-level implementation of machine learning systems, with experience in GPU programming and performance optimisation. Eager to contribute to impactful projects while continuously improving skills and knowledge.

Education

- **University of Nottingham** — Nottingham, UK
  Master of Science - Machine Learning in Science — *September 2023 - September 2024*
- **Royal Holloway, University of London** — Surrey, UK
  Bachelor of Science - Computer Science (Artificial Intelligence) — *September 2020 - July 2023*

Experience

- **PwC ETIC** — On-site
  AI Engineer — *June 2025 - Present*
  - Contributing to internal initiatives exploring agentic AI applications to enhance PwC client service workflows.
- **EcoMetric Ltd** — Hybrid
  Machine Learning Researcher (Placement) — *April 2024 - September 2024*
  - Developed a Proof of Concept (POC) framework using Pix2Pix cGAN to convert SAR imagery to optical (RGB), improving the visual interpretability of satellite data for agricultural fields and mitigating cloud cover issues.
  - Integrated multi-spectral image conversion techniques, extending model capabilities to generate spectrally more accurate imagery, enabling remote measurement of soil organic carbon and improving precision in environmental monitoring.

Publications & Technical Writing

- **Paged Attention from First Principles: A View Inside vLLM**: Wrote a detailed blog explaining the gap between LLM training and inference, covering prefill vs decoding, KV caching, and memory fragmentation. Broke down how paged KV caching and PagedAttention solve these at scale. Included an appendix on techniques like continuous batching, speculative decoding, and quantisation. Blog **(30+ bearblog up votes)**
- **AWQ: Activation-Aware Weight Quantisation**: Authored a deep-dive on quantising LLMs for edge devices. Discussed quantisation granularity (group, channel, and tensor-level), weight-only vs weight+activation schemes, derived symmetric quantisation, and explained how AWQ improves accuracy via weight–activation scaling. Blog

Projects

- **High Performance GEMM CUDA Kernel on NVIDIA H100 (Ongoing)**: Designing and implementing custom CUDA kernels for high-performance General Matrix Multiplication (GEMM), specifically optimised for NVIDIA H100 (Hopper) architecture. Focused on maximising compute throughput and memory efficiency by leveraging architecture specifics, including inspecting SASS/PTX assembly and optimising memory transactions. Rigorously benchmarked performance against cuBLAS achieving on par performance, detailing methodologies and results in an in-depth technical blog. (Ongoing) - **Skills:** GPU optimisation, CUDA, High-Performance Computing GitHub Blog
- **GPT-OSS-20B: PyTorch Implementation of a Mixture-of-Experts Transformer**: Developed a full PyTorch implementation of the GPT-OSS-20B architecture from scratch, focusing on efficient inference and modular extensibility. Implemented key components including RoPE with YaRN and NTK-by-parts scaling for long-context support, RMSNorm, SwiGLU, Mixture-of-Experts (MoE) with top-k routing, grouped-query attention (GQA) with learned sinks and banded attention. Authored detailed documentation for each component. **Skills:** PyTorch, Transformer Architectures, Mixture-of-Experts GitHub **(30+ GitHub stars)**
- **Attention-Enhanced CNNs vs Spectral-Spatial Transformers for Precise Crop Classification from UAV-borne Hyperspectral Images**: Designed and implemented deep learning models for precision agriculture, achieving 99.72% classification accuracy. Enhanced spectral-spatial transformers and attention-CNNs, improving baseline performance by 0.2–0.6%. Integrated 3D convolution kernels for in-network spectral dimension reduction, enhancing feature extraction. **Skills:** Computer Vision, Vision Transformers, CNNs, Python, PyTorch Paper GitHub
- **FlashAttention 2 Optimisation with Triton**: Implemented the forward pass of FlashAttention 2 using Triton to optimise memory efficiency and inference speed for Transformer models. Achieved up to 30× throughput improvement over a naive manual PyTorch baseline on large sequences, with comparable performance to PyTorch's scaled dot-product implementation. Benchmarked on an A100 using Triton's testing suite. (July 2025) - **Skills:** Triton, GPU programming GitHub
- **Convolutional Neural Networks for Scaled Autonomous Driving Using PiCar**: Trained CNNs for edge and object detection to navigate a Raspberry Pi Car, achieving real-time performance and placing 4th place in a Kaggle competition with a loss of 0.0127. Optimised ResNet and MobileNet models using TensorFlow Lite with quantisation for faster inference. (May 2024) - **Skills:** Computer Vision, Model Quantisation, TensorFlow, Keras, CNNs, Object Detection, Edge Detection, Python Paper GitHub

Skills Summary

- **Programming Languages**: Python, Java, C++
- **Frameworks & Libraries**: PyTorch, TensorFlow, Triton, CUDA, Hugging Face Transformers, Scikit-learn, NumPy, Pandas, SciPy, Matplotlib
- **Development Tools & Methodologies**: Git (Version Control), Test-Driven Development (TDD), Agile Methodologies
- **Languages**: English (Fluent), Arabic (Native)
- **Soft Skills**: Communication, Critical Thinking, Teamwork, Problem Solving, Time Management