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
X: https://x.com/hamzaelshafie

Professional Summary

Machine Learning Systems Engineer focused on inference, GPU programming, and LLM systems. Experienced in implementing and benchmarking custom low level ML kernels in CUDA C++ and ThunderKittens DSL. Strong background in PyTorch deep learning modelling, with 4+ years of Python experience and a Master's degree in Machine Learning. Write in depth technical blogs on GPU programming, inference engineering, and ML systems, with open source contributions to GPU kernel DSLs.

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 - June 2026*
  - 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

- **Dissecting ThunderKittens: Anatomy of a Compact DSL for High-Performance AI Kernels**: Wrote an in-depth technical deep dive on ThunderKittens, dissecting its core abstractions and how they map onto modern GPU hardware. Covered tile and vector abstractions, global layouts, warp and warpgroup compute, TMA, swizzling, Hopper WGMMA, Blackwell tcgen05, 2xSM MMA, tensor memory, Cluster Launch Control, pipeline templates, and persistent scheduling. Demonstrated TK's `lcf` pipeline template by implementing a non-causal attention prefill kernel and benchmarking it against FlashAttention-2 and FlashAttention-3 on an H100 PCIe, beating FA2 by around 1.55x on average in my setup and tracking FA3 closely at longer sequence lengths. The blog was added to the main ThunderKittens repository by Stanford's Hazy Research Lab. Blog
- **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 **(75+ 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 (70+ GitHub Stars)**: 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. **Skills:** GPU optimisation, CUDA, High-Performance Computing GitHub Blog
- **GPT-OSS-20B: PyTorch Implementation of a Mixture-of-Experts Transformer (200+ GitHub Stars)**: 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
- **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

- **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

Open Source Contributions

- **ThunderKittens**: Contributed a performance optimisation to register tile reductions by replacing the previous shuffle down plus leader broadcast pattern with XOR butterfly shuffles. Updated row and column reduction paths, inspected generated SASS, and validated the change on H100 GPU. PR

Skills Summary

- **Programming Languages**: Python, CUDA C++
- **GPU Programming**: CUDA, ThunderKittens DSL, Triton, PTX/SASS inspection, Nsight Compute, Nsight Systems
- **ML Systems**: LLM inference, vLLM, KV caching, prefix caching, KV cache offloading, P/D disaggregation, distributed serving, batching, quantisation, hardware-aware attention
- **Frameworks & Libraries**: PyTorch, NumPy
- **Development Tools**: Git, Linux, benchmarking, profiling
- **Soft Skills**: Communication, critical thinking, teamwork, problem solving, self-directed, persistent
- **Languages**: English (Fluent), Arabic (Native)