Oliver Grainge

LinkedIn: www.linkedin.com/in/oliver-grainge
GitHub: github.com/olivergrainge
Email: oliver@grainge.me
Mobile: +44 748 188 153

Professional Summary

ML Engineer with a strong research background in deploying and optimizing machine learning models. Expertise in quantization, distributed systems, and MLOps pipelines to reduce latency and memory usage while ensuring scalability in both cloud and edge environments.

Skills

- **Languages**: Python, C/C++, CUDA, Triton, SQL
- **Technologies**: SLURM, Git, pytest, pandas, PyTorch, TensorFlow, ONNX, TensorRT, FastAPI, ROS, Scikit-Learn, MLflow, Weights & Biases, OpenCV, Llama.cpp, Huggingface
- **Expertise**: Generative AI Deployment, LLMs, Computer Vision, CI/CD, Distributed Systems, Data Engineering

Experience

- **AI Security Institute** — Remote
  Named Researcher — *Nov 2024 – Current*
  - **Assessing VLM Geo-localization**: Developed a framework to evaluate the capabilities of proprietary and open-source vision-language models for geo-localization.
  - **Privacy Mitigation**: Designed mitigation techniques to address privacy risks in geo-localization without compromising system usability.
- **Arm** — Remote
  Contract Researcher — *Nov 2024 – Current*
  - **Edge AI Inference Optimization**: Optimized latency and memory usage on Cortex-A devices using internal SIMD intrinsics and multi-threading in C++ and Python.
  - **Cloud AI Inference Optimization**: Enhanced throughput on Neoverse Graviton servers by applying graph optimization, INT8/INT4 quantization, and pruning techniques.
- **Foster & Partners** — Southampton, UK
  Research Collaborator — *July 2019 – Oct 2019*
  - **Automated Quality Control**: Implemented a proof-of-concept mapping and localization system for construction quality control using OpenCV, ROS, and optimization in Python and C++.
- **Gartner** — London, UK
  Lead Delivery Consultant — *Aug 2019 – Mar 2020*
  - **Digital Banking**: Led UK digital banking market data collection using PowerBI and Excel, ensuring thorough data cleaning and effective client presentations.
  - **Digital Strategy**: Managed a team of 4 interns to gather and verify UK telecommunication market data, drafting initial client reports.

Projects

- **NeuroCompress Toolkit**: Developed an open-source Python library for training binary and ternary neural networks to facilitate efficient model deployment.
- **1.58bit Vision Transformer**: Pre-trained a 1.58-bit ViT on ImageNet-1k with custom Triton kernels, achieving a 57% memory saving and 23% latency reduction for a 4% accuracy drop.
- **Visual Navigation**: Implemented a complete 3D visual mapping and localization pipeline using OpenCV, enhancing real-time navigation capabilities.
- **Face Identification System**: Built an open-source client-server application for on-edge face identification in a home security system.

Education

- **(iPhD) Machine Intelligence** — Southampton, UK
  Thesis: Efficient Resource-Constrained Visual Place Recognition — *Oct 2022 – Current*
- **(BEng) Electronics and Electrical Engineering** — Southampton, UK
  Key Courses: Deep Learning (81), Optimisation (83), Machine Learning (76); Graduated 1st ClassSept 2019 – Jul 2022 (83%)

Publications and Awards

- **TeTRA-VPR: A Ternary Transformer for Compact Visual Place Recognition** — arXiv
  Developed a two-stage VPR training pipeline with progressive 2-bit quantization, reducing memory by 65% and latency by 35% without accuracy loss. — Mar 2025
- **Design Space Exploration of Low-Bit Quantized Visual Place Recognition** — IEEE RAL
  Established design rules for training and deploying visual AI embedding models on embedded devices using PyTorch, ONNX, and TensorRT. Achieved 63% latency reduction and 95% memory savings with a 0.5% R@1 decrease (Accepted). — Jun 2024
- **Structured Pruning for Efficient Visual Place Recognition** — IEEE RAL
  Developed architecture-specific pruning techniques reducing feature extraction latency by 21% and memory by 16% with under 1% R@1 drop (Accepted). — Aug 2024
- **Zepler Institute Award** — Southampton, UK
  Received the award for academic excellence with a dissertation score of 88%. — Jul 2022