PEDRO GIMENES

PhD Researcher in Machine Learning Systems

@ pgimenes@outlook.com +44 7757 025295 pedrogimenes.co.uk in /pedrocgimenes /pgimenes

INTERNSHIPS

RTL Design Intern (GPU Hardware Engineering)

Apple Inc.

April 2022 → September 2022 St Albans

- Implemented new hardware features to support extended requirements for **memory management** and **interrupt handling**.
- Used **Formal Verification** tools to accelerate feature bring up and minimize bugs ahead of RTL delivery. Also maintained **Sequential Equivalence Checking** and **Clock Domain Crossing** infrastructure.

Undergraduate Engineer (GPU Debug Infrastructure Team)

Arm Ltd.

July 2021 → March 2022 Cambridge

- Developed Python libraries aimed at parsing and visualization of simulation results to identify **top-level** bugs.
- Contributed to the development of a Model/Emulator GPU Testbench aimed at increasing visibility of FPGA Debug IP.

OTHER PROJECTS

MASE: Machine Learning System Exploration Tools

DeepWok Lab

- Machine Learning compiler for efficient inference deployment of language and vision models through a custom intermediate representation enabling **hardware and software co-optimization**.
- Implemented compiler passes for quantization, pruning and automatic parallelisation in distributed platforms.

AMPLE: Event-Driven Accelerator for Mixed-Arithmetic GNN Inference on Large Graphs (Master's Thesis)

DeepWok Lab

- Custom accelerator for Graph Neural Network (GNN) inference on large graphs, leading to a mean **speedup of $243\times$ and $7.2\times$** against CPU and GPU counterparts over a range of graph datasets.
- Introduced a novel **event-driven programming flow**, reducing pipeline gaps by addressing the non-uniform distribution in node degrees. Also developed a **mixed-arithmetic architecture**, enabling inference over graphs with nodes quantized at node granularity.

PUBLICATIONS

[1] Zeyu Cao et al. *Scaling Laws for Mixed quantization in Large Language Models*. 2024. arXiv: 2410.06722 [cs.CL].

EDUCATION

PhD in Deep Learning Systems

Imperial College London

2023 → 2027 London

Pursuing research in the following topics, supervised by Dr. Aaron Zhao and Dr. George Constantinides.

- Reasoning with Large Language Models
- Adaptive Inference Serving Systems
- Heterogeneous Distributed Systems
- Hardware-Aware Architecture Search
- Mixed-Precision Neural Networks

MEng in Electrical & Electronic Engineering (EEE)

Imperial College London

2019 → 2023 London

- Course average: **77.96%**
  US equivalent: 4.0 GPA.
- A Levels: A*AA in Mathematics, Further Mathematics and Physics from OCR.

AWARDS

- 2023: **Dean's List Award** for being placed in the Top 10% of the yearly cohort at Imperial College.
- 2022: **Dean's List Award** for being placed in the Top 10% of the yearly cohort at Imperial College.
- 2021: **IET Horizons Bursary**, awarded to students overcoming personal obstacles to pursue engineering education.

SKILLS

Programming Languages

Python C++ SystemVerilog

Libraries and Tools

PyTorch transformers vLLM PyTorch Geometric

Languages

Portuguese (native) English (fluent) Italian (B1) French (B1)