David Pissarra

david.pissarra@nyu.edu davidpissarra.com davidpissarra David Pissarra

Education

New York University, Courant *New York City, USA*

Ph.D. in Computer Science *2024-Present*

Co-advised by Prof. Aurojit Panda and Prof. Jinyang Li.

Instituto Superior Técnico, University of Lisbon *Lisbon, Portugal*

M.Sc. in Computer Science and Engineering *2022-2023*

Grade A, **GPA:** 19/20

M.Sc. Thesis Grade: 19/20

Tsinghua University *Beijing, China*

Graduate Visiting Student in Department of Computer Science *2021-2022*

Grade A, **GPA:** 3.93/4.00

The 1st year of the Master's was completed at Tsinghua University.

Instituto Superior Técnico, University of Lisbon *Lisbon, Portugal*

B.Sc. in Computer Science and Engineering *2018-2021*

Academic Merit Diploma, excellent performance

3-year B.Sc.

Research Interests

- Machine Learning Systems
- Compiler Optimization
- Large-scale Machine Learning
- Networking

Experience

Catalyst Research Group, Carnegie Mellon University *Pittsburgh, USA*

Research Intern *Sep 2023 - Dec 2023*

Worked on Machine Learning Compilation for Large Language Models.

Advised by Prof. Tianqi Chen.

Fraunhofer Portugal *Lisbon, Portugal*

Machine Learning Researcher *Jun 2023 - Jul 2024*

Worked on deploying and fine-tuning Large Language Models to perform internal tasks.

INESC-ID *Lisbon, Portugal*

Student Researcher *Sep 2022 - Sep 2023*

Master's Student Researcher. Advised by Prof. Fernando Ramos (IST Lisbon) and Prof. Muhammad Shahbaz (Purdue University).

Sensei Tech *Lisbon, Portugal*

Machine Learning Researcher *Summer 2022*

Designed a Machine Learning system capable of automatically detecting annotation errors in image retail product datasets. Image Dataset Anomaly Detection.

Deloitte *Lisbon, Portugal*

Software Engineer Intern *Summer 2021*

During this internship, contacted with various SAP technologies, completing an openSAP course addressing the SAP Cloud Application Programming Model (SAP CAP). Developed a small SAP CAP demo application, connecting it to multiple APIs and other SAP modules (such as SAP Fiori).

Publications

Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study *PrivateNLP @ ACL 2024*

David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, Vitor Rolla

Peregrine: ML-based Malicious Traffic Detection for Terabit Networks *preprint arXiv*

João Amado, Francisco Pereira, **David Pissarra**, Salvatore Signorello, Miguel Correia, Fernando Ramos

Teaching

Big Data and Machine Learning Systems *New York University*

Teaching Assistant *Fall 2025*

Grad course. I am currently TAing the Big Data and Machine Learning Systems course.

Artificial Intelligence *Instituto Superior Técnico*

Teaching Assistant *Spring 2024*

Undergrad course. I was invited to teach the lab component of the AI course.

Distributed Systems *Instituto Superior Técnico*

Teaching Assistant *Spring 2024*

Undergrad course. I was invited to teach the lab component of the Distributed Systems course.

Search and Planning *Instituto Superior Técnico*

Teaching Assistant *Fall 2023*

Grad course. Responsible for the course project, based on the Patient Transportation Problem (PTP), and also for project office hours.

Computer Organization *Instituto Superior Técnico*

Teaching Assistant *Fall 2021*

Undergrad course. Responsible for evaluating students, solving practical problems in classes, as well as aiding them during the development of their projects.

Selected Projects

Ongoing Project *Fall 2024 - Present*

Research Project

In this research project, I focus on analyzing hidden optimization opportunities in modern deep learning workflows which cannot be discovered by current ML compilers. Given the widespread adoption of deep neural networks (DNNs), my research goal is to design tools capable of generating efficient lower-level code for a given DNN architecture. Essentially, the main problem I am trying to tackle is to take coarse, high-level DNN computational graphs and automatically identify pieces of computation that can be accelerated for a target hardware. Broadly, using ML compiler optimization, there are two approaches I am targeting to automatically optimizing training and inference: *Advanced Tile Scheduling* and *Graph Super-Optimization*.

MLC LLM *Summer 2023 - Present*

Open-source Contributor

MLC LLM is a universal solution that allows any language model to be deployed natively on a diverse set of hardware backends and native applications, plus a productive framework for everyone to further optimize model performance for their own use cases. Everything runs locally with no server support and is accelerated with local GPUs on phones and laptops, by leveraging TVM. In this project I was responsible for implementing several LLM architectures on the framework, as well as multiple system optimizations, such as quantization and KV cache schemes (e.g., quantized KV Cache, sliding window KV Cache, paged KV Cache etc.), and also ensuring integration of multiple frameworks with MLC-LLM, e.g., CUTLASS.

In-Network ML-based Anomaly Detection *Spring 2023 - Fall 2023*

M.Sc. Thesis

We worked on a workflow to run Deep Neural Networks (DNN) inference in programmable switches, enabling detection at packet time scales. Running Network Intrusion Detection Systems (NIDS) in the switch is very limited to the device constraints. To address this problem we explored different DNN models, and some techniques to further optimize their execution, such as quantization and feature importance.

Honors & Awards

2024 **MacCracken Fellowship** *New York University*

2021 **B.Sc. Academic Merit Award** *Instituto Superior Técnico*