Franz Louis T. Cesista

ML Research Scientist at Expedock | 2x IOI and ICPC World Finalist | BS Mathematics at AdMU

Email | franzlouiscesista@gmail.com
Github | github.com/leloykun
LinkedIn | linkedin.com/in/franzcesista
Newsletter | ponder.substack.com

Work Experience

Machine Learning Research Scientist at Expedock July 2023 - Mar 2024

- *Robust & Cheap ML Infrastructure*: Trained, deployed, and setted-up monitoring for hundreds of small, multi-modal models using Nvidia's Triton Inference Server on AWS' Sagemaker Inference Platform. The system was able to handle 6,000 documents per day with a peak load of *250 per minute* (with ~100% uptime) on a single Nvidia A10G server.
- *Developed End-to-End ML Data Pipelines:* Constructed the data infrastructure for our entire AI pipeline, from data collection, preprocessing, streaming to model training, benchmarking, and the continuous monitoring of the models' performance.
- *Achieved SOTA Performance:* Led the research and development of a novel approach to augmenting LLMs for Business Document Information Extraction tasks such as Key-Information (Localization and) Extraction and Line Items Recognition. This approach was able to beat the current SOTA by up to 10% F1 score on public benchmarks and up to 33% on our internal benchmarks.

Full-Stack Software Engineer at Expedock June 2021 - July 2023

- *Reduced Distributed System Faults by ~99%:* Employed defensive engineering on our core data fetching and reconciliation processes with various Trade Management Systems such as Cargowise. Also simplified abstractions used in the logistics industry that emerged due to having a paper-based bureaucracy.
- *Productionized 11 Data Products for Logistics Management:* Collaborated closely with cross-functional teams to build the data pipeline and visualizations of the first set of interactive data products we sold to our customers. These data products help logistics executives make more data-driven decisions straight from the data we parse from the documents they send us.
- *Simplified UX for Data Query Building:* Developed an inhouse declarative framework that takes in configurations and produces a simple, flexible, but general-enough UX for data filtering and visualizations--ala Metabase, but more programmatically customizable.
- *Streamlined Data Analytics:* Developed internal engineering tools that automate tedious analytics work such as a tool that allows us to train a model on our data with only one line of code, run an inference job with another line of code, and run explainability tools with another line of code. This ultimately helped the product team decide which segments of the logistics industry to prioritize.

Data Science Intern at Exora - PH Summer, 2020

- Collaborated closely with the startup's leadership team to (1) strategize data collection and (2) train machine learning models for forecasting energy supply and demand in the Philippines.

Education

Ateneo de Manila University, *BS Mathematics* 2018 - 2021

- Team 1 member of the Competitive Programming Varsity team
- Co-Founder and Former CTO of Google Developer Student Clubs - Loyola Branch

Python • C++ • CUDA • PostgresQL • Snowflake • DBT • React • TypeScript • ReactRedux • GraphQL • PyTorch • Keras • Scikit-Learn • AWS SageMaker • Nvidia Triton Inference Server • Terraform • Docker

Awards and Honors

Competitive Programming

2021 - 2022	**ICPC World Finals Participant (2x)** Inter-Collegiate Programming Competition - Russia (2021) & Bangladesh (2022)
2018 - 2019	**IOI World Finals Participant (2x)** International Olympiad in Informatics - Iran (2018) & Japan (2019)
2016 - 2018	1 Silver & 2 Bronze Medals National Olympiad in Informatics - Singapore (Invitational)
2018 - 2020	Regional Finalist (2x) & Best Local Team & Bronze Medal International Collegiate Programming Contest - Singapore & Kuala Lumpur & Manila & Jakarta

Hackathons

2020	Top Philippine Team, Top 11 in the ASEAN Region *Shopee Codeleague* ● Competed with more than 20,000 participants on data science, data analytics, and algorithmic challenges. ● Notable tasks: Unsupervised machine translation & fine-tuning pretrained large multilingual language models such as XLM-Roberta.
2018	Champion *Hack4PH - Digitizing Public Information Category* ● Built OpenAPI, an API for querying publicly available government data. This led to high-level discussions on a potential buy-out.

Side Projects

Expedock AutoML	Expedock's AutoML Library. Train a model on data from Snowflake with just *one line of code* and run predictions on another line of code.
Flash Attention Minimal	A minimal implementation of Flash Attention 1 & 2 in just ~300 lines of CUDA code. This is still a work-in-progress, but the ultimate goal is to implement the various variations of Hyperbolic Attention in CUDA.
LLama2.cpp	A C++ implementation of Meta's Llama2 generative large-language model. I also optimized the original C implementation by Karpathy by adding parallelization on the multi-head attention layer.
ICPC Team Notebook	[Lead Maintainer] Ateneo programming varsity's code library. Used by everyone in the organization as reference material for coding competitions.
AI for Sea	Geotemporal booking demand prediction for Grab's AI for SEA Challenge.
Codeball AI	A rule-based AI built to beat other AIs in a Rocket-League-esque 3D soccer game. It managed to get to the finals of the Russian AI Cup 2018.
Codewars AI	A particle swarm-based AI built to beat other AIs in a Command-and-Conquer-esque game. It managed to get to the finals of the Russian AI Cup 2017.
Fact Police	An endpoint for a Messenger bot that can detect computer-generated (fake) text.