Jeffery Cao Siming

jeffery2@andrew.cmu.edu | (412) 214-2031 | linkedin.com/in/jeffery-cao-siming/

Education

Carnegie Mellon University (CMU) – School of Computer Science **Pittsburgh, PA**
Master of Computational Data Science: GPA: **4.04/4.33** *May 2023*
Selected Coursework: Distributed Systems, Cloud Computing, Machine Learning in Production, Deep Learning Systems, Multi-Modal Machine Learning, Intro to Deep Learning (A+), Machine Learning (PhD), Introduction to Computer Systems

Nanyang Technological University (NTU), School of Electrical and Electronic Engineering (EEE) **Singapore**
Bachelor of Engineering: Cumulative GPA: **4.88/5.00** *Jun 2021*
- Awarded Dean's List 2018/19 and 2020/21 (Top 5% in the cohort)
- Awarded Andy Grove Scholarship for Intel's Employees' Children
- Semester Exchange Program at the University of Waterloo, Canada (Sep 2019 – Dec 2019)

Skills

Software Development: Linux, Docker, Kubernetes, Terraform, AWS, Azure, Jenkins, Git, React.JS, Express.JS, Django
Machine Learning/ Big Data: Apache Spark, Databricks, Hbase, PyTorch, Tensorflow, Numpy, Scikit Learn, Pandas
Programming Languages: Python, C/C++, Go, Javascript, Java, Scala, SQL, GraphQL

Professional Experiences

Celonis Inc. New York City, NY
Software Engineer (Machine Learning) *Sep 2023 – Present*
- Developed an Automatic Prompt Generation module to help business users configure LLM prompts for Celonis, reducing time to value. The product was highlighted in Celosphere 2024.
- Authored a SIGMOD Demonstration Paper on Prompt Editor, a taxonomy-driven system to assist in the prompt writing process
- Implemented a LLM-based solution for automatically unblocking Credit Blocks that secured a seven-figure deal.
- Built a Data Governance system for Celonis' Copilot to ensure data quality for Retrieval Augmented Generation (RAG). The product was highlighted in Gartner Magic Quadrant for its contribution to product performance and quality.

Pinterest Palo Alto, CA
Machine Learning Intern *May 2022 – Aug 2022*
- Built a Location-based Recommender System enabling a new product feature that allows users to explore local interests
- Prototyped a Contrastive Learning method with BERT encoders to learn meaningful embeddings of geo-location
- Developed a production-ready pipeline in *Apache Spark* and automated the pipeline using a workflow (Airflow)

NDR Medical Technology Singapore
Machine Learning Engineer Intern *Jan 2021 – Jun 2021*
- Developed a Path Planning module (Python) for company's surgical robot, enabling it to avoid lung vessels during operation
- Implemented a modified *U-Net segmentation network* using Tensorflow to locate blood vessels in 3D CT Scans
- Initiated and built a manual path planner for surgeons to test their estimates, thus improving usability

Safemode Tel Aviv, Israel
Data Science Intern (NTU Overseas Entrepreneurship Program) *Jan 2020 – Apr 2020*
- Modeled driver's behavior with Gradient-boosted Trees, enabling the company to receive a 1.6 million shekels grant
- Developed an Extract, Transform and Load (ETL) pipeline on AWS cloud services to integrate partner's telematics data
- Utilized a spatially indexed data structure to classify 40million geospatial coordinates, decreasing runtime by a factor of >400

Projects

AI Guide Dog
- Developed the PredRNN spatiotemporal model for egocentric navigation, improving overall accuracy by 10% over the baseline
- Implemented Focal Loss to deal with an imbalanced dataset and increased validation performance by 2%
- Built a pipeline that model explanations using the GradCam technique, thus making the system more interpretable

Needle Framework
- Implemented an Auto-Differentiation library in Python and C++ that enables training of Neural Networks like PyTorch
- Developed Nvidia CUDA kernels in C++ to enable GPU support and parallelism, thus increasing matrix operation speed by ~10x
- Integrated SIMD and OpenBLAS instructions to optimize matrix multiplication, resulting in a 1.2x speedup over Numpy

Twitter Recommendation System (Spark ETL)
- Collaborated in a team of 3 to build a Tweet Recommendation System microservice from scratch using Go, Fasthttp web framework
- Developed ETL pipelines in *Apache Spark* to enable automatic transformation of data from the twitter API
- Deployed the service on AWS using Docker/Kubernetes/Terraform ensuring the service scales automatically with request volume