Lawrence Du

Machine Learning

larrydu88@gmail.com | 626-808-7096 | github.com/LarsDu | linkedin.com/in/LarsDu | dublog.net

2911 McKinley Dr Santa Clara, CA 95051

Skills

Techniques

Neural networks (Transformers, CNNs, GraphNN, Diffusion), Large Language Models (LLMs), Louvain/Leiden algorithm, SVMs, PCA, KNN, decision trees

Tools

PyTorch, TensorFlow, Numpy, Numba, Pandas, Sklearn, Flask, AWS, Google Cloud, Terraform, Pulumi, Docker, Metaflow, Kubernetes

Languages

Python, C/C++, Rust, SQL, C#, Java, Dart, Bash, HTML/CSS, Some Mandarin and Spanish

Experience

Freenome • ***Senior Machine Learning Research Engineer***

08/2022 - 06/2024 (South San Francisco, CA)

- Led greenfield project building end-to-end scalable distributed machine learning platform using PyTorch, Ray, and Kubernetes for cancer detection from deep sequencing (methylated DNA) and protein data, enabling training of much larger models leveraging data distributed parallel (DDP) processing speeding up model training by >10x.
- Deployed and managed an organization-wide MLFlow based model tracking system using Terraform, Pulumi, and Google Cloud enabling live-monitoring of deep learning model training progress, instantaneous results sharing, and completely automated and reproducible report generation - reducing researcher manual effort by at least 5x.
- Built scalable multitask learning, elastic net, and neural network based models in PyTorch with improved performance for classifying Colorectal Cancer risk from cell-free DNA data for a clinical trial cohort of >27,000 individuals.
- Piloted a project to summarize biomedical literature using LLMs, first using GPT-4 and later by fine-tuning an open source LLM via DPO (direct preference optimization), demonstrating the viability of using LLMs to parse unstructured biomedical records for scaling up feature extraction.

23andMe • ***Machine Learning Engineer*** • ***Data Scientist (prior to 2020)***

11/2018 - 08/2022 (Sunnyvale, CA)

- Created and deployed into production Recent Ancestor Locations (RAL) - a high precision, high recall country matching algorithm which serves >15 million customers worldwide.
- Improved graph-based techniques for unsupervised identification of populations by genetically based identity-by-descent (IBD) family relationship, demonstrating an effective way to segment sub-populations (graph community detection) in Mexico and the United Kingdom in an semi-unsupervised manner.
- Built a large-scale feature engineering ETL pipeline for imputed SNPs (~10 million samples x ~1 million SNPs) using AWS Batch, Metaflow, AWS Glue, and AWS Athena enabling creation of higher quality GWAS and Polygenic Risk Score (PRS) ML models.
- Developed improved models for type 2 diabetes and Coronary Artery Disease by building and evaluating model stacking ensembles into production PRS pipelines, improving the sensitivity and specificity of 23andMe tests for tens of thousands of customers.
- Automated performance metric report generation for all polygenic risk score classifiers leveraging MLFlow artifact storage and headless Jupyter execution, reducing researcher time spent on analysis from days to minutes.

Scripps Research • ***Bioinformatician IV***

05/2018 - 10/2018 (San Diego, CA)

- Developed a classifier for organ transplant rejection using RNA data and wrote pipelines for Nanopore long-read sequencers using Common Workflow Language.

Juno Diagnostics • ***Independent Consultant***

09/2017 - 02/2018 (San Diego, CA)

- Developed patent – US20210020314A1 - Deep learning-based methods, devices, and systems for prenatal testing along with a Tensorflow based classifier for detecting prenatal genetic abnormalities from high throughput sequencing data.

Insight • ***Data Science Fellow***

01/2017 - 04/2017 (Remote)

- Built and deployed (as a Flask app on AWS EC2) DeepPixelMonster - a Tensorflow based GAN for creating pixel art, back when GANs were still relatively state-of-the art.

UC San Diego • ***PhD Student Biology*** • ***Scott A. Rifkin Lab***

08/2010 - 05/2017 (La Jolla, CA)

- Wrote DeepNuc - a CNN model for classifying over 500,000 transcriptional start site (TSS) flanking sequences from humans, mice, fruit flies, and nematodes as well as for over 60,000 microRNA target sequences.
- Researched the role of RNA expression noise during animal development by imaging single molecule RNA expression data in >5,000 embryos and analyzing data using self-written MATLAB tools for image segmentation, fluorescence quantification, and image deconvolution.

Education

Ph.D Biology UC San Diego, 2010 - 2017

B.A. Biological Sciences *Genetics and Development, Magna Cum Laude* Cornell University, 2006 - 2010

Activities and interests

DuBlog (https://dublog.net) • Diffumon - Simple DDPM image generator • Developing the VR game Rogue Stargun (https://roguestargun.com) • Ludum Dare Game Jams • Blender3D