YUFAN ZHANG

New York, NY | 914-720-6892 | yufanbruce@gmail.com | GitHub | LinkedIn | Website

EDUCATION

CORNELL TECH (CORNELL UNIVERSITY), New York, NY — Aug 2023 – May 2025
M.S. in Information Systems | GPA: 4.0/4.0 | Merit Scholarship Recipient

DUKE UNIVERSITY / DUKE KUNSHAN UNIVERSITY, Suzhou, China — Aug 2019 – May 2023
B.S. in Data Science | GPA: 3.7/4.0 | Dean's Lists

TECHNICAL SKILLS

Coding Languages: Python (PySpark, Pandas, NumPy, Matplotlib, Seaborn), SQL, Java, Matlab, R, HTML/CSS/JavaScript
Machine Learning & AI: PyTorch (Lightning), TensorFlow, Scikit-Learn, HuggingFace, OpenCV, NLTK, LangChain, Ray, Wandb
Miscellaneous Skills: AWS (SageMaker, EMR), Azure (AI Studio), GCP (BigQuery), Git, Docker, Jupyter Notebooks

EXPERIENCE

INSITRO | Machine Learning Engineer Intern, South San Francisco, CA — May 2024 – Aug 2024

- Onboarded a **ViT**-based cell segmentation model, boosting AP@0.5 from **0.79** to **0.86** by integrating a **DETR** object detector to generate bounding boxes for prompting a **Segment Anything** Model in mask creation.
- Migrated the previously deployed segmentation model, *Cellpose*, to **PyTorch Lightning**, enabling multi-GPU training on **Azure AI studio**.
- Monitored and organized over **100+** training jobs using Weights & Biases (**Wandb**), ensuring efficient tracking and reproducibility of experiments.
- Developed unified helper functions for seamless data access across **AWS S3** and **Azure Blob Storage**, facilitating the company's migration in computing infrastructure from AWS to Azure, now used by **5+** teams to streamline model training workflows.

EBAY | Product Manager Intern, Shanghai, China — Mar 2023 – Jun 2023

- Executed improvements to eBay's internal data streaming platform by leveraging cloud computing tools (**Apache Kafka, Flink**), facilitating the integration of **10** additional use cases in 3 months for improved functionality and a **13%** user satisfaction score improvement.
- Executed in-depth technical research to support the platform's deployment design of Kafka for high availability across data centers, which has supported **100+** clusters across **3** data centers and processed over a **trillion** data streams daily.
- Conducted data analysis and visualization with **PowerBI** on user feedback surveys, deriving actionable insights for the engineering team, leading to 4 new data storage connectors in the platform, enhancing the platform's data accessibility.

DUKE KUNSHAN UNIVERSITY | Research Assistant, Suzhou, China — Jul 2022 – Nov 2022

- Designed and implemented an end-to-end **GAN**-based generative model with **PyTorch** for cross-language font image style transfer, resulting in a first-authored paper at **ACM Multimedia 2022**, the top multimedia computing conference, with **5 citations**. [Paper] [GitHub]
- Integrated a **self-attention** mechanism into the image style encoder to effectively capture both local and global font styles, enhancing the model's ability to generalize across diverse font styles, leading to a **14%** improvement in ablation studies.
- Developed an adaptive **skip connection** mechanism to improve content fidelity, leading to a **12%** improvement in SSIM in ablation studies.

PROJECTS

Retrieval Augmented Generation (RAG) LLM Application for Ray Documentation, (Python) — [GitHub] Fall 2024

- Developed a RAG-based LLM application using **Python**, **Ray**, **Langchain**, and **PostgreSQL**, enabling efficient question-answering on Ray documentation with retrieval latency of less than **100ms** and improved response accuracy by **24%** compared to not using RAG.
- Implemented and evaluated advanced RAG techniques (multi-query, HyDE), achieving a **13.2%** improvement in retrieval quality.
- Built an evaluation pipeline using **GPT-4** as a scoring agent, used to optimize retrieval and generation configurations (chunk size).

Data-drive Restaurant Recommendation System for Yelp, (Python, Spark) — [GitHub] [Report] Spring 2024

- Engineered restaurant recommendation systems with **PySpark**, achieving an RMSE of **1.081** using a hybrid recommendation approach, which is a **13.5%** improvement over the content-based filtering approach and a **51.7%** improvement over the ALS-based collaborative filtering method.
- Conducted data cleaning and **feature engineering** on 2 million users and 150,000 businesses data with **Python** and **Pandas**, including scaling numerical attributes, one-hot encoding for categorical attributes, and ordinal encoding with custom scoring functions.

Text-to-SQL Translation by Training, Fine-tuning and Prompting LLM, (Python, PyTorch) — [GitHub] [Report] Spring 2024

- Experimented with 3 NLP techniques to build **domain-specific LLMs**, including training from scratch, fine-tuning and prompt engineering.
- Fine-tuned a T5 language model using **Hugging Face** on the text-SQL pair data, achieving the best F1 score of **0.627** over other approaches.
- Implemented **few-shot prompting** for LLMs on text-to-SQL, leading a **37.3%** F1 score improvement over the baseline prompting method.

miniTorch: Python Re-implementation of the Torch API, (Python) — [GitHub] Fall 2023

- Engineered an alternative library to the **Torch** API with **Python** and **Numba**, resulting in **100%** compatibility with native **PyTorch** code.
- Architected a custom Tensor data structure pivotal for deep learning model training and evaluation, supporting tensor backend operations including **broadcasting**, mathematical operation **overloads**, **auto-differentiation**, and **backpropagation**.