Furkan Ozyurt

+1 (618) 802-0464 furknozyurt@outlook.com New York, United States

EXPERIENCE

Data Scientist (Contract)

Amgen

03/2022 06/2023 Cambridge, United States

- Fine-tuned large language models (e.g., BERT, ROBERTA, GPT 3) on company documents to perform text summarization, generation, paraphrasing, classification, and question answering.
- Developed and deployed a deep learning pipeline to predict whether changes in documents need to be reported. Integrated the deployed model into an application and implemented performance monitoring. The pipeline reduced decision-making time **by up to 95%.**
- Built and deployed a RAG pipeline that extracted and returned the needed information from the documents. This reduced information search time **by up to 99%**.
- Developed robust, scalable, and automated ETL/data pipelines to provide the team with reliable, high-quality, and up-to-date data.
- Optimized a previously developed machine learning pipeline and reduced runtime from 18 hours to 4 hours (a **75% reduction**) using Spark.

Associate Engineer (Contract)

Amgen

09/2020 03/2022 Cambridge, United States

- Developed a deep learning pipeline that classified documents into right categories and integrated it into an application which saved the department approximately **$500,000.**
- Designed and implemented an end-to-end machine learning pipeline in AWS SageMaker to forecast product consumption for key company products across multiple locations.

PROJECTS

Self-Learning

11/2023 02/2024

https://github.com/ozyurtf/self-learning

- Worked on a project with the goal of developing self-learning system that enables a simulated truck to back up to a target position from any initial location autonomously without collecting any data manually.
- Developed a custom loss function customized for the challenges of the task because standard loss functions were not useful.
- Built two separate models: one to create internal representation of the environment in which the truck operates and another to determine the optimal steering angle for the truck's next move based on the internal representation of the environment.
- Successfully trained these models to enable the truck to consistently reach the target position smoothly no matter where it is initialized.

Attention in CUDA

03/2025 - present

https://github.com/ozyurtf/attention-cuda

- Implementing multi-head attention mechanism in CUDA by utilizing shared memory, coalesced memory, warp shuffle, and tiling.
- Profiling and optimizing CUDA kernels using Nsight Systems and Nsight Compute to reduce inference latency.

EDUCATION

Master of Science - Computer Science

New York University - Courant

09/2023 05/2025 New York, United States

Bachelor of Science - Industrial Engineering (Mathematics Minor)

Istanbul Technical University

08/2016 05/2020 Istanbul, Turkey

- (Ranked in the **top 0.9%** of students in the national university exam)
- (Jointly completed the program with Southern Illinois University - Edwardsville)

SUMMARY

Master of Science student in Computer Science at New York University, with 3 years of industry experience in building data pipelines and training/optimizing/deploying monitoring machine learning and deep learning models. Possesses strong theoretical knowledge of various deep learning architectures (e.g., CNNs, RNNs, LSTMs, Transformers, Autoencoders, VAES, GANS, and Diffusion Models). Has a strong background in GPU architecture and CUDA.

KEY ACHIEVEMENTS

Decision-Making Time Reduction

Reduced the decision-making time by up to **95%** with a deep learning pipeline that predicts whether changes in documents need to be reported.

Information Search Time Reduction

Built and deployed a RAG pipeline that extracted and returned the required information from the input. This reduced information search time by up to **99%**.

Document Classification Savings

Built and deployed a deep learning document classifier. This saved the department **$500,000.**

Pipeline Runtime Optimization

Reduced the runtime of a machine learning pipeline by **75%** using Spark's parallel processing capabilities.

SKILLS

Programming Languages & Query Languages
Python, SQL, C/C++, CUDA

Machine Learning & AI
Machine Learning, Deep Learning, Natural Language Processing (NLP), Large Language Models (LLMs), Retrieval Augmented Generation (RAG), MLOPs

ML Frameworks & Libraries
PyTorch, Tensorflow, MLFlow, Huggingface

Data Engineering & Big Data
Databases, Databricks, Delta Lakes, Spark

Cloud Computing
AWS (EC2, S3, IAM, Athena, EMR, Glue, Redshift, Sagemaker), Microsoft Azure

Development Tools
Git, Docker, Kubernetes

FIND ME ONLINE

Personal Website
https://ozyurtf.github.io/

Github
https://github.com/ozyurtf

Linkedin
http://www.linkedin.com/in/ozyurtf/