WEI ZHANG

weizhang45name@gmail.com | (917)-330-6286 | www.linkedin.com/in/weizhang45 | www.weizhang45.com

EDUCATION

COLUMBIA UNIVERSITY New York, NY

Master of Science in Computer Science (Machine Learning Track) Feb 2025

- Relevant Coursework: High Performance Machine Learning, Advanced Software Engineering, Computer Vision I, Operating Systems, Cloud Infrasture/Computing, Applied/Theory Machine Learning, Artificial Intelligence, Database, Natural Language Processing, Data Structures (C/C++).

PROJECTS

High Performance Machine Learning: Optimizing Transformer Inference Speed Oct 2024 – Dec 2024

- Fine-tuned Pretrained Transformer models (Bert, GPT2, T5) on text classification dataset (MRPC, COLA), achieving accuracy ≥85%.
- Accelerated inference using post-training dynamic quantization and unstructured pruning (2x speedup).
- Enhanced efficiency of GPT-2 with Scaled Dot Product Attention (SDPA) and Flash Attention 2, achieving 1.5x inference speedup.

Cloud Computing: CUThen Networking/Chat App on AWS Sep 2023 - Dec 2023

- Developed a cloud-based networking/chat app on AWS for Columbia University students with similar background and hobbies, fostering meaningful connections.
- Hosted HTML/JavaScript frontend on AWS S3, integrated seamlessly with API Gateway and decoupled Lambda Functions, enabling flexible code updates and easier maintenance.
- Implemented Lambda Functions for profile management, real-time chat features, and matching algorithms, delivering a smooth and personalized user experience.
- Integrated Opensearch and DynamoDB for scalable user and chat data storage, and automated testing and deployment with AWS CodePipeline, achieving a fully automated workflow.

Machine Learning: Movie Scoring Project Jan 2023 - May 2023

- Built a movie scoring system by integrating 5 movie datasets (Film TV, Netflix, Amazon Prime, Hulu, and Disney+), processing information for 40,000+ movies.
- Designed regression models (linear regression with Lasso/Ridge) to predict movie rating based on 40,000+ movies with average mean square error of 1.2.
- Implemented classification models (Logistic regression, decision tree, KNN, and random forest) to determine if movie is better than average with accuracy of 80%.

WORK EXPERIENCE

Seamus Henchy and Associates, Inc New York, NY

Associate / IT Assistant Jun 2021 - Jan 2023

- Streamlined documentation processes by automating daily report generation, meeting agendas, and project record-documents using Python and Microsoft Flow, increasing efficiency by 30%.
- Collaborated with NYC architects and leading contractors to successfully construct school rooftop addition, and school basement renovation, meeting strict budget constraints and tight deadlines.

LANGUAGE AND IT SKILLS

- Computer Language/Skills: Machine Learning, Python, SQL Database, HTML/Javascript, C/C++, Java, Linux, scikit-learn, pandas, numpy, TensorFlow, Pytorch, Keras, AWS, Docker, Kubernetes, CI/CD, React, node.js, Ruby on rails, Unix, Git, Postman, Shell scripting, CUDA, predictive models.
- Concepts: Algorithm Design, Data Structures, Software Development Life Cycle(SDLC), Agile, Web Services, IoT, Saas, DevOps, Data science.
- Certifications: AWS Certified Solutions Architect - Associate, AWS Certified Cloud Practitioner.