Hooman Ramezani

hooman125@gmail.com | www.hoomanramezani.com | github.com/HoomanRamezani

Education

University of Toronto Toronto, ON
MASc in Machine Learning & Operations Research *2023 – 2024*

- Thesis: Building healthcare LLM and ViT models with multimodal clinical data for lung cancer treatment
- GPA: 4.0 / 4.0 — Coursework: Large Language Models, Cloud Data (Spark, AWS), Deep Learning Theory

University of Waterloo Waterloo, ON
BASc in Systems Design Engineering *2018 – 2023*

- GPA: 3.7 / 4.0 — Coursework: Intro. Deep Learning, Intro. Machine Learning, Pattern Recognition, Neuroscience

Experience

Machine Learning Researcher May 2023 – Present
University Health Network *Full Time – Toronto, ON*

- Authored 'Lung-DETR' Transformer architecture for lung cancer segmentation, achieving DiceC of 94.2% (SOTA)
- Tuned **multi-modal LLM** for treatment planning with 84% concordance, utilizing CT imagery and clinical notes
- Overcame class imbalance with 5% tumor rate with Detection Transformer, Focal Loss and precise data processing

Machine Learning Engineer Sept. 2022 – April 2023
Advanced Micro Devices (AMD) *Intern – Toronto, ON*

- Accelerated **LLM inference** on AMD CPUs by over 80%, applying pruning, quantization, knowledge distillation
- Built content recommendation model with candidate retrieval and reranking stages with 80% latency reduction
- Managed distributed training workflows to handle over 5 terabytes of cloud datasets within Spark and Azure

Machine Learning Engineer Jan. 2022 – April 2022
Apple (formerly DarwinAI) *Intern – Remote*

- Delivered 95% sensitivity CNN model to Pfizer for Liver Fibrosis diagnosis, reducing examination time by 40%
- Implemented MLOps CI/CD pipelines for model training, testing, and deployment accelerating delivery
- Collaborated with three customers to translate requirements into scalable deep learning solutions on edge devices

Machine Learning Engineer May 2021 – Aug. 2021
Applied Brain Research *Intern – Waterloo, ON*

- Achieved 94% object detection accuracy of defects with CNN model deployed on drone for real-time scanning
- Designed Unreal Engine 4 simulation to synthetically generate 50,000 annotated images to overcome lack of data

Research // Projects

Livy Education Chatbot ML Lead | *GPT-4, RAG, LangChain, Chroma* Jan 2024 – Present

- Leading AI for an assignment research chatbot for Canvas students, integrating GPT-4 and RAG
- Utilized Chroma vector database and fine-tuned embeddings for course material reranking, improving relevancy

Advanced Detection of Parkinson's Freezing of Gait | *CNN, Time-Series Data* 2023

- Achieved 94% accuracy for fall forecasting for early detection of Freezing of Gait (FOG) in Parkinson's patients
- Analyzed biometric data (EMG, ECG) utilizing novel time-series InceptionTime CNN model and data processing

Enhanced Robotic Grasping with PointNet | *LiDAR, Robotics, 3D Computer Vision* 2022

- Trained grasp proposition neurel-net achieving 87% grasp success rate, adapting custom PointNet deep learning models to identify optimal grasp points from LiDAR camera

Technical Skills

Languages: Python (NumPy, pandas, Matplotlib) , C/C++, Java, SQL, R, JavaScript
Frameworks: TensorFlow, PyTorch, Flask, RESTful APIs, Azure AI Studio, Bedrock, LangChain, Docker, Kubernetes
Libraries: Scikit-learn, Hugging Face, OpenCV, Keras, Retrieval Augmented Generation (RAG), CUDA
Cloud/Tools: AWS (S3, EC2, Lambda, SageMaker), Spark, Git