Dongim Lee

1000 Olin Way, Needham, Massachusetts • +1-781-809-4715 • dlee3@olin.edu • linkedin.com/in/dongim • dongim04.github.io/

Education

OLIN COLLEGE OF ENGINEERING *Needham, MA*
BS, Engineering Computing *Dec 2026*
• Relevant Coursework: Machine Learning, Reinforcement Learning Reading Group, Modeling & Simulation, HVAC Optimization
• GPA: 3.93/4.0

GLOBAL VISION CHRISTIAN SCHOOL *Republic of Korea*
High School Diploma *Jan 2023*
• Graduated as Valedictorian; 100% Tuition Merit Scholarship
• GPA: 4.29/4.30

Experience

ROAD SAFETY INTELLIGENCE WITH AUGMENTED LLM | MIT Break Through Tech, Michelin Mobility Intelligence *Mar 2024 - Present*
• Developing a natural language interface for geospatial analysis of LA crash data, utilizing LangChain for function calling, to handle complex queries within the geospatial dataset.
• Performed exploratory data analysis (EDA) and integrated Points of Interest data with crash data using Haversine formula for accurate distance calculations, to enable automated spatial data extraction and analysis using LLMs.
• Developing a responsive interface using Streamlit, integrating users' real-time GPS, to provide robust and interactive geospatial insights.
• Selected from 3k+ applicants for an AI program hosted by MIT, with hands-on Machine Learning coursework and industry mentorship.

FINE-TUNING ASR MODELS ON STUTTERING RECORDINGS (TEAM LEAD) | Olin Public Interest Technology *Sep 2024 - Present*
• Leading a team to address bias in Automatic Speech Recognition (ASR) models against stuttered speech using LibriSpeech/Stutter data.
• Identified disparities in Word Error Rate (WER), with OpenAI's Whisper showing a 2x increase and Facebook's Wav2Vec a 6.4x increase in transcribing stuttered speech; Successfully reduced Whisper's WER by 3% and Wav2Vec's by 33% by removing repeated words.
• Fine-tuned Wav2Vec on AWS SageMaker, built word tokenizers for Chinese stuttered speech, and uploaded the model to Hugging Face.

BENCHMARKING STUTTERING RECORDINGS AGAINST ASR MODELS | Boston University, AImpower.org *Jun 2024 - Jul 2024*
• Evaluated leading ASR models (Whisper, Google Speech-to-Text, Wav2Vec, Azure, WeNet) to assess bias in recognizing Mandarin stuttered speech.
• Segmented 50+ hours of Mandarin speech data with labeled transcriptions using Pydub, processed it on BU's Shared Computing Cluster, addressed hallucinations, and calculated WER, CER, BLEU (NLTK), WordNet Wu-Palmer Similarity, and GloVe Cosine Similarity.
• Demonstrated that more stutter segments lead to higher error rates, with WeNet achieving a WER of 0.30, outperforming Wav2Vec's 0.52; analyzed model performance across stutter types, revealing a 0.2 WER difference between sound repetitions and interjections.
• Authored comprehensive reports explaining model performance and bias analysis, including technical visualizations; Conducted weekly meetings and presentations with the client company's CEO.

POLITICAL DISCOURSE ANALYSIS ON SOCIAL MEDIA | Wellesley College Credibility Lab *Sep 2024 - Present*
• Designing a system to query political keywords on social media and analyzing how political discourse spreads across different platforms.
• Developed an automated pipeline to collect real-time trending political keywords and queries from Google Trends and upload them to an SQL database for efficient data storage and management.

INVERTED PENDULUM ROBOT SIMULATION TEAM | Olin Autonomous Robot Training Lab *Jan 2024 - May 2024*
• Developed a custom Gym environment to train an inverted pendulum robot to balance in the upright position, featuring real-time simulation visualization.
• Applied the Proximal Policy Optimization reinforcement learning (RL) algorithm, performing hyperparameter sweeps (e.g., learning rate, reward function, entropy coefficient) to optimize model performance; Used Weights & Biases for machine learning experiment tracking.
• Successfully solved the CartPole swing-up problem, achieving stable balance in simulation through iterative RL training.

WEEDER ROBOT WEED IDENTIFICATION TEAM| Olin Human-centered AI Research Lab, Farm Robotics Challenge *Jan 2024 - Apr 2024*
• Implemented OpenCV library to develop an autonomous weeding robot that minimizes soil inversion and preserves soil carbon.
• Employed DBSCAN for clustering, integrated PlantNet API for weed classification, and utilized ROS for camera integration.

Projects

AI-GENERATED IMAGE CLASSIFICATION MODEL *Oct 2024*
• Developed a deep learning model using CNNs to classify AI-generated images from MidJourney against real images, with 11 classes.
• Applied data augmentation techniques to enhance generalization, with max-pooling layers to prevent overfitting.
• Achieved up to 80% accuracy on test data, demonstrating the model's capability to distinguish between real and AI-generated images.

AI CALLIGRAPHY GENERATOR *Jul 2024 - Present*
• Building a calligraphy generator model to create personalized English calligraphy, merging deep learning with digital typography.
• Implemented the Vector Quantized Generative Adversarial Network and Contrastive Language–Image Pre-training (VQGAN+CLIP) algorithm, to generate high-quality images from text prompts.
• Training the model with OCR for automated labeling of calligraphy images, leveraging AWS S3 for data management and storage.

UNDP SUDAN 2024 CONFLICT EVENTS ANALYSIS *Oct 2024*
• Analyzed and visualized the relationship between conflict events, refugees movements, and food insecurity in Sudan from 2019 to 2024, using H3 hexagonal indexing for geospatial data analysis, identifying conflict hotspots, trends, and humanitarian impacts across regions.
• Calculated confusion matrices to reveal strong correlations between the variables. (e.g., r=0.85 between conflict events and food insecurity levels by region indicates that areas with higher conflict tend to experience more severe food insecurity.)

Certificates

MACHINE LEARNING FOUNDATIONS, Cornell University *Jul 2024*
DEEP LEARNING WITH PYTORCH: GENERATIVE ADVERSARIAL NETWORK, Coursera Course Certificates *Jun 2024*
MACHINE LEARNING WITH PYTHON, IBM *Dec 2023*

Skills

PROGRAMMING: Python • C/C++ • R • MATLAB • HTML
MACHINE LEARNING: PyTorch • TensorFlow • NLTK • OpenCV • Scikit-learn
DEVOPS & TOOLS: AWS • Docker • VMware • PostgreSQL • Linux/Unix • Web Scraping • Django
HARDWARE: Arduino • KiCad • SOLIDWORKS