XINGJIAN DONG

xdong404@usc.edu | LinkedIn | GitHub | dannydxj.github.io

SUMMARY

I am a Computer Science Master's student at USC, planning to apply for a PhD program in the near future. I am interested in broad topics about **Machine Learning** and **Natural Language Processing**. Currently, I am interested in how data from different stages of training influence the model's performance and behavior. I am working on **out-of-distribution and long-tail data detection and analysis** for large-scale pre-training data of LLMs.

EDUCATION

University of Southern California   **Aug. 2023 – Jun. 2025 (expected)**
Master of Science in Computer Science   *Los Angeles, USA*
- **GPA**: **4.0**/4.0

University of California, Berkeley   **Jan. 2023 – May. 2023**
Exchange Student in Computer Science and Artificial Intelligence   *Berkeley, USA*
- **GPA**: **4.0**/4.0
- **Courses**: Natural Language Processing, Undergraduate Research Apprentice Program (URAP)

Harbin Institute of Technology   **Aug. 2019 – Jun. 2023**
Bachelor of Engineering in Artificial Intelligence, Honor Graduate   *China*
- **Rank**: 1st/30
- **GPA**: **92.70**/100, **3.96**/4.0 (WES Certification)
- **Courses**: Data Structure, Algorithm, Operating System, Database System, Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, Speech Signal Processing, Intelligent Computing System

PUBLICATION

Out-of-Distribution Detection through Soft Clustering with Non-Negative Kernel Regression
*Aryan Gulati, **Xingjian Dong**, Carlos Hurtado, Sarath Shekkizhar, Swabha Swayamdipta, Antonio Ortega*
- EMNLP 2024 Findings

RESEARCH EXPERIENCE

USC DILL Lab   **Jan. 2024 - Now**
Research Assistant, advised by Prof. Swabha Swayamdipta   *Los Angeles, USA*
- Worked on applying EC-NNK Means on Anomaly Detection in language data, enhancing both speed and accuracy.
- Currently working on a project to detect and analyze the long-tail data in large-scale pre-training datasets for LLMs.

Berkeley Speech & Computation Lab   **Feb. 2023 – May 2023**
Undergraduate Research Apprentice, advised by Prof. Gašper Beguš   *Berkeley, USA*
- Engaged in the Articulation GAN project, employing GANs to create articulatory representations from stochastic noise for speech synthesis; innovated and empirically validated enhancements such as enabling full-parameter fine-tuning and introducing a diffusion model to do the generation.

Senior Thesis: QA System Based on BERT with Syntax Embedding   **Feb. 2023 – Jun. 2023**
Advised by Prof. Yongdong Xu   *China*
- Proposed a novel approach, *SEBERT* (<u>BERT</u> with <u>S</u>yntax <u>E</u>mbedding), which utilizes a Transformer syntax parser for syntax representation extraction, transforms these into Syntax Embeddings, and integrates them with BERT Embeddings; this innovation maintains the original BERT's performance level on QA tasks.
- Constructed an interactive website demo using Gradio to showcase the results of various models including the Transformer syntax parser, original BERT QA model, multiple SEBERT QA models, QA model with GPT-3.5 API, and QA model with GPT-3.5 API and KNN-based few-shot prompting.

Academic Activities

HUAWEI Ascend All-wisdom Project: MindX Pre-trained Model   **Sept. 2021 – Dec. 2023**
In Charge of MobileNetV3-Small   *China*

- **Model Development**: Enhanced and re-trained the model using MindSpore, ensuring compatibility with the inference requirements of the Ascend AI Processor.
- **Testing and Validation**: Conducted rigorous model tests on ImageNet, which included a training test on the ModelArts AI Platform and two offline inference tests on Ascend 310 using the MindX SDK and MxBase module. Verified that Top1 and Top5 accuracies, as well as model functions, met the specified criteria.
- **Outcomes**: Crafted detailed test reports and user documents; earned the title of **"Ascend All-wisdom Developer"**; the model has been **incorporated into the HUAWEI Ascend open source community**.

ROBOMASTER Robotics Competition   **Sept. 2020 – Oct. 2021**
Member of Vision Department of HIT Combat Robot Team, HERO   *China*

- Engineered a high-performance "self-aiming system" that combined digital image processing for armor extraction, CNN for armor confirmation, and MCU data communication via CAN protocol, resulting in a 99%+ accuracy in the actual competition field. [code]
- Designed a soft-triggered "timing system" that precisely synchronizes images with IMU pose data using industrial camera SDK and multi-threading operations, reaching a frame rate of 100+ fps.
- Undertook on-site debugging and maintenance of the Sentinel, a fully autonomous robot, during the competition stage; the robot achieved notable success, including a double kill in a single game and the winning of the Combat Award for its damage impact.[code]
- **Additional Projects and Achievements**: Introduced an advanced "four-point" model for the self-aiming system based on an improved YOLOX-Nano, applied a PnP algorithm for monocular ranging, and developed a "mineral localization system" for the Engineer, a functional robot, using a CNN-based model; clinched the **Second Prize** in RoboMaster University Championship (RMUC) Final Tournament 2021, ranking in the **top 5%** among 300+ teams.

Selected Awards

Honor Graduate of HIT | 10% of all graduates	**Jun. 2023**
HUAWEI *Intelligent Base* Scholarship | 1% of 1k+ students	**Oct. 2022**
First-Class People's Scholarship | 1/30, 2 times	**Oct. 2020 & Jun. 2022**
Outstanding Student | 10% of 400+ students, 2 times	**May 2020 & Dec. 2021**
Technological Innovation Scholarship	**Nov. 2021**
National Second Prize in *RoboMaster University Championship (RMUC)* | 5% of 300+ teams	**Aug. 2021**
Social Work Scholarship	**Jun. 2021**
Outstanding Student Leader	**Dec. 2020**
Second Prize in *National Mathematical Contest in Modeling*	**Oct. 2020**

Technical Skills

Languages: Python, C/C++, Java, SQL, Go.
Tools: Git, Linux, LaTeX, OpenCV, ROS, CMake, Industrial Camera
Deep Learning Frameworks: PyTorch, TensorFlow, MindSpore