Dongyu Yao

+86 13971862248 ◊ Email ◊ Homepage ◊ Wuhan, Hubei, China

EDUCATION

Wuhan University (WHU) **Bachelor of Engineering in Cyberspace Security** 09/2020-06/2024

- Junior GPA: 3.98/4, Major GPA: 3.83/4, Overall GPA: 3.80/4 (90.1/100), Ranking: **4/123** (Top 3.25%)
- **Scholarships and Honors**: National Scholarship (0.2% national-wide), Pacemaker to Merit Student (0.1% school-wide), First Class Scholarship (5% school-wide)

PUBLICATION & PATENT

(Paper) Dongyu Yao and Boheng Li, *Dual-level Interaction for Domain Adaptive Semantic Segmentation*

- Accepted by ICCV Workshop on Uncertainty Quantification for Computer Vision (UnCV), 2023
- Reviewing results from *ICME 23'*: 1 strong accept (5), 1 weak accept (4), 1 weak reject (2), **Average (3.67/5)**

(Paper) Dongyu Yao, Jianshu Zhang , Ian G. Harris, andMarcel Carlsson, *FuzzLLM: A Novel and Universal Fuzzing Framework for Proactively Discovering Jailbreak Vulnerabilities in Large Language Models*

- Submitted to International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024 for reviewing

(Patent) A Method and a System of Dual-level Interacted Domain Adaptive Semantic Segmentation

- First inventor, Patent application ID: 2023102210107

RESEARCH EXPERIENCES

⋆ Research Assistant at Aerospace Information Security and Trusted Computing (AI Sec) lab (Wuhan University)

Supervised by Prof. Run Wang 2022/07-Present

(Artificial Intelligence - Computer Vision) Domain Adaptive Semantic Segmentation with Self-Training Method

- Proposed a dual-level learning framework for unsupervised domain adaptation (UDA) in semantic segmentation and other self-trained UDA scenarios to enhance the adaptation of segmentation models on unlabeled target domain data
- Devised a labeled instance bank to tackle the storage issue when incorporating instance information
- Introduced the dual-level pseudo-label interaction (first time in such tasks) to improve the accuracy and robustness of pseudo-labels
- Outperformed the state-of-the-art models by **2.7%** and **2.5%** on two widely used benchmarks (especially on confusing and long-tailed classes)
- **A Paper accepted by ICCV Workshop on Uncertainty Quantification for Computer Vision (UnCV), 2023**

⋆ Research Assistant at UCInspire Summer Research Program of University of California Irvine

Supervised by Prof. Ian G. Harris 2023/06-2023/09

(AI Security - Security in NLP) A Fuzzing Framework to uncover Jailbreak Vulnerabilities of Large Language Models

- Leveraged the traditional fuzzing technique to uncover jailbreak vulnerabilities in LLMs, aiming to craft sufficient jailbreak prompts to ensure both syntactic and semantic variation while maintaining the structure of each prompt
- Generalized 3 base classes of jailbreaks capable of merging into more powerful combo attacks
- Decomposed a prompt into 3 basic components that could be combined with each other to craft numerous test samples
- Conducted extensive experiments regarding fuzzing tests on 6 open-sourced and 2 commercial LLMs.
- Proved the effectiveness and comprehensiveness of our testing framework through comparisons with existing jailbreak prompts and jailbreaking methods
- **Submitted a paper to International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024**

⋆ Research Assistant at Network Information System Security & Privacy (NIS&P) Lab (Wuhan University)

Supervised by Prof. Qian Wang 2022/03-2022/08

(AI Security - Biometric Authentication) FreeGait: Gait-Recognition-based Identification and Authentication System

- Replicated the cell phone sensor-based gait recognition method and mastered the deep learning framework
- Discussed the application of "Gait ID" and its feasibility and security for future commercial promotion
- Developed corresponding gait data acquisition software and gait authentication server based on previous learning
- Coded the CNN-LSTM architecture for gait recognition with Python and TensorFlow, and debugged the server for better data transfer

- Achieved a 90% recognition accuracy on a 500-people dataset and successfully authenticated users' identities
- Related project has been open-sourced, and won the First Prize in WHU Information Security Competition

SELECTED PROJECTS

Team Project: **Advoice: Proactively Defend against AI-Synthesized Fake Voices via Adversarial Attacks** 03/2023-06/2023

- Utilized adversarial examples to disrupt the voice synthesis models, thus proactively protecting the voice of a victim
- Added the psycho-acoustic model to increase the imperceptibility of adding adversarial perturbations into the victim's voice sample
- Changed the original defense mode from end-to-end to encoder-only, increasing the defense framework's transferability

Independent Project: **Rain's Search Engine - A customized Search Engine** 04/2023-05/2023

- Developed a customized command line search engine (shown as a webpage) with multiple input processing techniques form scratch to index 37600+ TDT3 news articles (code is open-sourced at Github)
- Used Python and Whoosh Library to build index, implemented the function for query analysis, and enabled four types for searching, including search for matched words, matched texts, ambiguity search and stemming search
- Coded the document relevance score function (TF-IDF score) independently, using HTML and Flask library to code a UI web page and results visualization

Team Project, **Anthropomorphic IP Creation and Cultural Derivative Design** 12/2022-05/2023

National Second Prize *(10% nation-wide) in Computer Design Competition (Digital Media Technology - Static Design Track)*

- Introduced anthropomorphic IP of the "Compendium of Materia Medica" to promote Chinese medicine
- Established the homepage for the IP based on GitHub Pages, using PHP and JavaScript to design a user-friendly interface for seamlessly interactions with markdown files

Team Project: **Entity-Stance Analysis (Dataset Construction and Model Fine-tuning)** 12/2022-02/2023

- Implemented sequence labeling techniques by incorporating stance information into the tags to recognize entities (Vaccines) and their stance in textual input
- Appended the stance label directly to the original tags considering the characteristics of data types, creatively transforming the task into "Stance-Embedded Entity Recognition"
- Collected and annotated the VaxESA dataset containing 3,000 articles on vaccine stances from various media websites
- Sourced and deployed 3 Named Entity Recognition (NER) models from tier-1 NLP conferences as baseline models and conducted further fine-tuning and evaluations on these models
- Employed Python and HTML for the visualization and deployment of the model

TEACHING EXPERIENCES

Teaching Assistant of Big Data Analysis, Undergraduate Course, Wuhan University 09/2023-01/2024

- Assigned and graded the course projects of students, presented three English lectures regarding "data efficient scene understanding"

Teaching Assistant of Artificial Intelligence, Undergraduate Course, Wuhan University 09/2022-01/2023

- Conducted class Q&A, managed and graded lab assignments, and led the end-of-term review sessions

EXTRA-CURRICULAR ACTIVITIES

WHU Symphony Orchestra Chief Percussionist & Drummer of CandyAfterX (16-year Playing Exp.) 09/2020-Present

- Played the timpani and the snare drum in several concerts, including the 2021 New Year's Concert at Qintai Concert Hall of Wuhan and the 22's Midsummer Music Festival at Wuhan University, invited Livehouse performance at Rainbow Pub
- Participated in the Sixth National University Art Exhibition and won the Second Prize (15% nationwide)

ADDITIONAL INFORMATION

Programming: C, C++, Python, MATLAB, HTML, Markdown

Libraries: Numpy, Pandas, PyTorch, OpenCV, Tensorflow

Developer Tools: VS Code, PyCharm, Jupyter Notebook, LATEX, Git

English Proficiency: TOEFL 112 (Speaking & Writing 28), GRE 330 (Writing 4.0)