Dongyu Yao

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

ACADEMIC BACKGROUND

Wuhan University (WHU) Bachelor of Engineering in Cyberspace Security **2020/9-2024/6 (expected)**

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

PUBLICATION & PATENT

(Paper) Dongyu Yao, 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, Marcel 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

- When applied to an unknown scenario, the performance of segmentation model trained on a single domain dataset drops severely. Meanwhile, it is extremely expensive to annotate data of different scenes. Thus people leverage much cheaper synthetic data as source domain to train a model and then adapt it to the target domain. However, the gap between different domains reduces the efficiency
- Unsupervised Domain Adaptation (UDA) is explored to generalize the network trained with labeled source (synthetic) data to unlabeled target (real) data. Even though Current advances have mitigated noisy pseudo-labels resulting from the domain gap. However, they still struggle with erroneous pseudo-labels near the boundaries of the semantic classifier
- (Our method) Proposed a Dual-level Learning framework for UDA in Semantic Segmentation and other self-training UDA scenarios
- (Our method) Devised a labeled instance bank to tackle the storage issue when incorporating instance information; Introduced the dual-level pseudo-label interaction (first time in the task) to improve the accuracy and robustness of pseudo-labels
- Outperformed the state-of-the-art by **2.7%** and **2.5%** on two widely used benchmarks (especially on confusing and long-tailed classes)
- **Related paper has been 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

- Jailbreak Prompt -- The human-engineered prompt to circumvent the safety measures, and induce content that violates service guidelines. When first discussed online in March, 2023, this vulnerability has captured the attention of research community
- Recent defense and detection methods against this vulnerability are very passive, model owners need to wait for an attack to be identified as effective. Current research only explores the semantic variants of prompts, which lacks diversity in the broader category of jailbreaks
- (Our method) we proposed to leverage the traditional fuzzing technique to uncover jailbreak vulnerabilities in LLMs. The key objective is to craft sufficient jailbreak prompts to ensure both syntactic and semantic variation while maintaining the structure of each prompt
- (Our method) For an automatic fuzzing framework, we generalized three base classes of jailbreaks that can merge into more powerful combo attacks. We decomposed a prompt into three basic components that can combine with each other to craft numerous test samples
- We conducted extensive experiments regarding fuzzing tests on 6 open-sourced and 2 commercial LLMs. The comparison with existing jailbreak prompts and jailbreaking methods proves the effectiveness and comprehensiveness of our testing framework
- **Related paper has been submitted 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

- The authentication system of biometric has been popular among research and industry communities. However, some biometrics require obtrusive authentication process (such as fingerprints) or they are easy to forge and attack (such as facial images and human voices).
- In comparison, the gait biometric is hard to conceal and has the advantage of being unobtrusive. Inertial sensors, such as accelerometers and gyroscopes, are commonly integrated into smartphones, which makes gait data convenient and inexpensive to collect.
- (Our work) Our team 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

PROJECT EXPERIENCES

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

Content Security Final Project

- Voice synthesize systems using deep learning techniques have become overwhelmingly powerful and accessible in recent years. Given the speech of the target voice, such a system e.g. voice conversion (VC), text-to-speech (TTS), can generate synthetic fake speeches that speak in the voice of the target speaker, which can be natural and realistic, and thus are exhibiting real threats to society

- Unfortunately, existing works that propose to defend against forgeries mainly focus on developing detectors for passively detecting whether a given speech is real or fake, which not only struggle to catch the rapid progress of forgery techniques in this cat-and-mouse game but also failed to block the spread of misinformation in advance
- We proposed to utilize 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. This helps increase the transferability of our defense framework
- Graded as the **Best Course Project** (97 out of 100)

Rain's Search Engine -- A customized search engine from scratch **Independent Project** **2023/04-2023/05**

Information Retrieval Final Project

- Requirement: Write from scratch a customized command line search engine to index 37,600+ TDT3 news articles
- Result: A complete search Engine (shown as a web page) with multiple input processing techniques. Code is open-sourced at Github
- Use Python and Whoosh Library to build index. Implement the function for query analysis, enables **four types** (three more than required) of searching, including search for matched words, matched texts, ambiguity search and stemming search. Code the document relevance score function (TF-IDF score) by myself. Use HTML and Flask library to code a UI web page and results visualization
- Graded as the **Best Course Project** (96 out of 100)

Entity-Stance Analysis (Dataset construction and model fine-tuning) **Team Project** **2022/12-2023/02**

Public Opinion Analysis Final Project

- Requirement: Recognize entities and their stance in textual input, focusing on "Cross-Strait Relations" and "Vaccines" topics
- Implemented sequence labeling techniques for the task, with an innovative design of sequence tags. Traditionally, sequence tags encapsulate only entity boundaries and category information. In this project, we seamlessly incorporated stance information into these tags. Given that stance, boundary, and type information do not overlap with each other, we appended the stance label directly to the original tags, innovatively transforming the task into "Stance-Embedded Entity Recognition".
- Collected and annotated the VaxESA dataset, containing 3,000 articles on vaccine stances sourced from various media websites
- Sourced and deployed three Named Entity Recognition (NER) models from tier-1 NLP conferences as baseline models. Conducted further fine-tuning and evaluations on these models. Employed Python and HTML for the visualization and deployment of the model
- Graded as the **Best Course Project** (92 out of 100, Class Average Score: 81)

The Secret of Accessing websites **Team Project** **2023/03-2023/06**

First Prize (5% division-wide) in Computer Design Competition (Computer Basics Tutorial Track)

- Starting with "Uncovering the Secrets Behind Websites", a concise and easy-to-understand approach was adopted to explain the inner workings of websites. Key concepts were elucidated using vivid and relatable analogies, steering clear of overly specialized or cumbersome terminology, ensuring even beginners can effortlessly grasp the content. Our video is available here.
- We abstracted principles of network communication into everyday scenarios. We broke down the communication process into multiple small steps. Utilizing multimedia tools such as images, animations, and videos, the teaching material was connected to real-world application scenarios. Through hands-on exercises, students were allowed to experience the network communication process firsthand

Strategy on Return Maximization **Team Project** **2022/01-2022/02**

Honorable Mention (15% worldwide), MCM/ICM Mathematical Contest in Modeling

- Established an XGBoost regression model with the stride of 5 to forecast the next day's price of Bitcoin and achieved a forecast accuracy of 80%; Implemented a dynamic programming strategy with MATLAB for optimal decision-making on trading behaviors including purchasing, selling, and holding to maximize returns;
- Composed an English paper for the contest regarding this project

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

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

- With an intent to breathe new life into traditional Chinese medicine promotion, we introduced anthropomorphized IP of the "Compendium of Materia Medica." At a time when most promotions feel a bit "same-old", our distinctive and endearing medicine characters offer a memorable take, effectively simplifying the conveyance of traditional Chinese medicine themes.
- On the technical front, we've efficiently leveraged GitHub Pages for the swift and cost-effective deployment of our IP's homepage. Using PHP and JavaScript, we established a user-friendly interface that seamlessly interacts with markdown source files. A straightforward navigation bar ensures users can swiftly access various content modules.

TEACHING EXPERIENCES

Teaching Assistant of Artificial Intelligence, Undergraduate Course, Wuhan University **Fall 2022**

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

Teaching Assistant of Big Data Analysis, Undergraduate Course, Wuhan University **Fall 2023**

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

EXTRACURRICULAR ACTIVITIES

WHU Symphony Orchestra Chief Percussionist & Drummer of CandyAfterX (16 years of playing experience) **2020/09-Present**

- Played the timpani and the snare drum in several concerts, including the 2021 New Year's Concert at Qintai Concert Hall of Wuhan City 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)

Invited talk (by New Oriental Education) **2023/04**

- Independent lecture on "How to learn TOEFL listening and speaking efficiently", 100+ audience, 2000+ blog views

ADDITIONAL INFORMATION

English Proficiency: TOEFL 112 (Speaking & Writing 28), GRE 324 (Writing 4.0)
Programming skills: C, C++, Python, MATLAB, HTML, Markdown
Developer Tools: VS Code, PyCharm, Jupyter Notebook, LATEX, Git
Libraries: Numpy, Pandas, PyTorch, OpenCV, Tensorflow