Lang Gao

E-mail: gaolang1643@outlook.com / Phone: +1 574-333-1750

Address: No. 68, Zhangzhuang Road, Huaiyin District, Jinan, Shandong, China, 250000

Education Background

Huazhong University of Science and Technology (HUST)   **Wuhan, China**

School of Computer Science & Technology   Sept. 2021 – Jun. 2025

Bachelor of Engineering in Computer Science and Technology

GPA: 4.34/5.0

Publications

Liu, Y.*, Gao, L*., Yang, M*., Xie, Y., Chen, P., Zhang, X., & Chen, W. (2024). VulDetectBench: Evaluating the Deep Capability of Vulnerability Detection with Large Language Models. arXiv preprint arXiv:2406.07595.

Xie, Y*., Zhou, C*., Gao, L*., Wu, J*., Li, X., Zhou, H.-Y., Liu, S., Xing, L., Zou, J., Xie, C., & Zhou, Y. (2024). MedTrinity-25M: A large-scale multimodal dataset with multi-granular annotations for medicine. NeurIPS 2024 Datasets and Benchmarks Track Submission.

Liu, W., Deng, Z., Niu, Z., Gao, L., Wang, J., Wang, H., & Li, R. (2024). Attacking for inspection and instruction: The risk of spurious correlations in even clean datasets. In Proceedings of the Neural Information Processing Systems Conference (NeurIPS).

Research Experiences

MedTrinity-25M: A Large-scale Multimodal Dataset with Multi-Granular Annotations for Medicine

First Co-Author, Instructed by Prof. Yuyin Zhou of University of California, Santa Cruz   Jan. 2024 – Jun. 2024

"A comprehensive, large-scale multimodal dataset for medicine, covering over 25M images across 10 modalities, with multi-granular annotations for more than 65 diseases. "   【**Link**】

➢ **Image Data Collection and Pre-processing**
● Collected 104 original image packages from sites like Kaggle, TCIA, and Zenodo, and kept 85 packages from them to ensure a balanced distribution, resulting in an image dataset of 25M samples
● Constructed an image transformation toolkit with libraries of nibabel, SimpleITK, and OpenCV, to achieve lossless image format conversion, normalization, and segmentations

➢ **Data Management**
● Deployed the annotation model, LLaVA-Med++, along with data shards to be annotated, on different servers for improving data generation speed, and merged data back together after annotation
● Monitor and adjust data to ensure that the distribution of modes and physiological structures is as uniform as possible

➢ **Design of the Multi-granular Annotation Pipeline**
● Designed a pipeline to combine image annotation, global and local features of the image, and external medical knowledge; and developed the prototype of the pipeline
● Applied an assembly of specialized medical grounding models and segmentation models to annotate local features in images(drawing ROIs), each used in one specific modality(e.g. dermatoscopy, fundus, chest X-rays, etc.), and used a visual language model for consolidating information from images lacking textual descriptions into text, resulting in the realization of multi-granular image annotations

VulDetectBench: Evaluating the Deep Capability of Vulnerability Detection with Large Language Models

First Co-Author, Instructed by Prof. Wei Chen of HUST   Jan. 2024 – Jun. 2024

"A novel, comprehensive benchmark, specifically designed to assess the code vulnerability detection capabilities of LLMs"   【**Link**】

● Obtained code vulnerability data from specific data sources such as NIST, Devign, and Big-Vul, and referred to the Common Weakness Enumeration (CWE) to construct a tree-structure vulnerability relationship graph, to ensure that the CWE type distribution in the benchmark is balanced and has similar levels
● Designed the benchmark as a combination of 5 progressively difficult tasks to observe the performance and capabilities of LLMs in code vulnerability analysis from multiple perspectives
● Designed novel metrics for distinguishing performance code analysis across different tasks from different aspects(the first to consider the open analysis of LLMs in vulnerability detection tasks)
● Tested 17 LLMs(both open-source and close-source) using VulDetectBench, and analyzed their performances. We concluded that currently, LLMs fall short on specific, more detailed vulnerability analysis while they perform well on simple tasks such as classification

LLM-Based Chinese Academic-Text Grammar Error Detection and Correction System Development

Group Research Member, Instructed by Prof. Ruixuan Li of IDC-Lab, HUST   May. 2023 – Apr. 2024

"An LLM-based Chinese 'Grammarly', especially focusing on scientific research scenario."

- Collected and curated initial scientific Chinese grammar correction dataset by crawling over 9,000 Chinese academic papers, and applied CLG-CGEC (a linguistic rules-based faulty sentence generation tool) to construct text correction pairs
- Integrated other relevant open-source benchmarks to evaluate LLM's performance on text correction comprehensively
- Applied few-shot learning and in-context learning techniques to stabilize the output of the ChatGLM2-6B model, to generate corrected sentences effectively
- Used p-tuning for efficient parameter fine-tuning, and took evaluation indicators (such as M2-Scorer) to evaluate model performance
- Data augmentation and enhancement: utilized GPT to expand incorrect sentences and applied NLP tools to do NER and sentence component substitution to obtain more complex text correction samples
- Introduced the more powerful InternLM2-13B model to support new features, including text polishing and paper evaluation report generation

Attacking for Inspection and Instruction(A2I): Recognizing Model-added Spurious Correlations for Faithful Explanation

Group Research Member, Instructed by Prof. Ruixuan Li of IDC-Lab, HUST   Oct. 2023 – Dec. 2023

"An interpretable causal model framework employing adversarial learning, aimed at correctly learning useful information from data with spurious correlations."

- Within the A2I code framework, a common classic self-explanatory model architectures were implemented using RNN as the carrier: Rationalizing Neural Predictions (RNP).
- Integrate a noise injection algorithm(NI), which enhances the robustness of the RNP architecture, into the RNP framework as a comparative experiment within A2I.
- Adjusted hyper-parameters to keep the results' gap within a controllable range (2%) during experiments, and conducted broad tests on interpretability benchmarks.
- The RNP algorithm is applied to graph neural networks for further exploration.

Awards & Honors

National Third Prize, The 5th Global Campus Artificial Intelligence Algorithm Elite Competition,2023.

National Third Prize, iFlytek Developer Competition, NLP Track, 2023

National Second Prize, The 5th Integrated Circuit EDA Design Elite Challenge (Deep Learning Track), 2023

National Second Price,15th China College Students' Service Outsourcing Innovation and Entrepreneurship Competition，2024

Optics Valley Morning Star Scholarship, Hubei Province, China, 2023

Scholarship for Academic Excellence, Huazhong University of Science and Technology, 2022

Skills

Computer Science Professional Skills & Tools:

Deep Learning Framework: Proficient in **Pytorch, Tensorflow**

Large Language Models:

    Proficient in Prompt Engineering(Chain-of-Thought,In-Context-Learning and few-shot learning)

    Fine-tune techniques(PEFT,full-parameter training;large-scale distributed training on server cluster),

    Transformers

Strong Data Management and Processing Skills: deduplication, cleaning, formatting, and statical analysis.

Programming Languages: Proficient in **Python, Linux, C, C++**

Language Skills: Chinese (native), English (advanced)