Gaoxiang Huang

gaoxianghuang@outlook.com · (+86) 132-8598-3328 · Personal Web

Summary

Experience in Explainable Artificial Intelligence research area and Multi-Modal Models image edit research area of Computer Science. In-depth knowledge of standford cs231n, D2I Mu Li and HKUST AIAA5023 et al. Skilled in research tools like Python, Pytorch, git and shell. Enthusiastic, Entrepreneurial, Highly-motivated. Work well both independently and as part of a team.

Education

The HongKong University of Science and Technology (Guangzhou) 2024.09 – Present

- Master of Philosophy Majoring in Artificial Intelligence
- CGA: 3.323/4
- Research interests: Explainable AI, Multi-modal learning, transfer learning
- Research advisor: Prof. Yutao Yue

China University of Petroleum Beijing 2020.09 – 2024.06

- Bachelor of Science Majoring in Statistics
- GPA: 80.9/100, 21.3%
- Research interests: System simulation, Social network analysis, Statistics learning

Publications

- Gaoxiang Huang, Songning Lai, Yutao Yue. Toward a more holistic Interpretability: A lightweight disentangled concept bottleneck model. (Under reivew)
- Songning Lai, Yu Huang, Jiayu Yang, Gaoxiang Huang.et al. Guarding the Gate: ConceptGuard Battles Concept-Level Backdoors in Concept Bottleneck Models. (Under review)

Scientific Competition

- *First Prize* China Undergraduate Mathematical Contest in Modeling(CUMCM) Sep. 2022
- *Grand Prize(3/4983) "Renzheng Cup"* National Student Mathematical Modeling Competition Jun. 2022
- *H Prize* Interdisciplinary Contest In Modeling(ICM) Feb. 2022
- *Third Prize "Huajiao Cup"*National University Mathematics Competition Nov. 2021
- *Second Prize* China Student Computer Design Competition Jun. 2022
- *Second Prize "MathorCup"* National Student Mathematical Modeling Competition Aug. 2022
- *Second Prize* Nation University math network championship Sep. 2022
- *H Prize*, *"Shuwei Cup"* International Student Mathematical Modeling Competition Nov. 2021

Honors and Awards

- Technology and Innovation Scholarship (China Univ. of Petroleum Beijing)
- Second-class Scholarship (China Univ. of Petroleum Beijing)
- HKUST(GZ) RedBird scholarship(240,000 Yuan)

Extracurricular Activity

- The 6th Shenzhen Mountain and Sea Hiking Challenge: 32km traverse of Dapeng Peninsula with an elevation gain of 1134 meters.
- Supported Hiking Traverse of Hong Kong's MacLehose Trail (Sections 2 to 9)

Research Experience

JD – Algorithm Intern Jul. 2025 – Oct. 2025

- Legal Assistant Agent (end-to-end post-training process of legal vertical large model: CL, SFT, RLHF). Initially, manual review consistency was 0.44 and case recall rate was 0.33; after optimization, consistency increased to 0.96 and recall rate to 0.93.
  - Participated in end-to-end post-training (CL/SFT/RLHF) of Legal Assistant Agent vertical model, raising manual review consistency (0.44→0.96) and case recall rate (0.33→0.93).
  - CL: Developed enterprise-level data desensitization tool (concurrency >30) processing 53k+ long texts in 48h; continued pre-training deepseek-r1 671B on 8-GPU H200 (0.73/0.69).
  - SFT: Generated SFT data via prompt engineering; proposed anchor set-aided generation; LoRA fine-tuning lifted to 0.91/0.81.
  - RLHF: Built DPO data tool based on user feedback, resolving inconsistencies to reach final indicators.

LLM-Driven Conscious and Memory-Enabled Companion Digital Human (Agent)
Guangzhou Nov. 2024 – Present

- Agent Module: Adopted Qwen3-14B as the pre-trained model and conducted single-GPU fine-tuning (A800 80G) on the CBT-bench dataset.
- Interpretability Module:
  - Explored potential relationships between concept annotations in the dataset using the Apriori algorithm for association rule learning to enhance the interpretability of causal variables.
  - Designed an innovative framework of Latent Disentanglement Concept Bottleneck Models (LDCBMs). Compared with previous models on CUB, AwA2, and CelebA datasets, it achieved SOTA performance with a concept alignment rate improvement of 0.1242 ± 0.0576 and a label accuracy increase of 0.1316 ± 0.0203, enabling end-to-end interpretability from input to concept to output.

Backdoor Attack Defense for Concept Bottleneck Models Sep 2024 – Dec 2024

- Project Summary: In explainable AI (XAI), Concept Bottleneck Models (CBMs) enhance interpretability via understandable underlying concepts, but are vulnerable to concept-level backdoor attacks (hidden triggers in concepts causing undetectable misbehavior). First proposed Conceptguard defense: constructed poisoned datasets, divided data subsets, and used majority voting to mitigate data-driven backdoor impacts.
- Key Contributions:
  - Theoretically proved a minimum trigger size threshold, above which Conceptguard effectively defends against attacks (average backdoor success rate reduced by 30%).
  - Led CBMs baseline and Conceptguard experiments on CUB dataset, demonstrating improved concept accuracy. Identified cluster mechanism as an unsupervised method to avoid concept-level category conflicts and enhance feature learning for better concept correlation capture.

Economic Linkage Evaluation Based on Exponential Random Graphs and Cooperative Game Theory Macau Research Assistant Jun. 2022 – Dec. 2022

- Addressed the insufficiency of current global waste trade network description methods. Pointed out that existing exponential random graph methods struggle to accurately reflect annual trade network relationships amid global productivity growth. Utilized the R package tergm to construct a global waste trade network, reducing the complexity of same-year network analysis and studying its structural evolution and determinants. Research findings provided references for formulating enterprise carbon subsidy policies and data support for optimizing trade policies.

Improved Social Force Model Enhancing Psychological and Behavioral Heterogeneity
Beijing Jun. 2023 – Dec. 2023

- Proposed an enhanced system dynamics model integrating agent-based stampede risk assessment, entropy method, expert weights, and psychological force parameters. Simulation tests showed that the model improved stampede risk assessment accuracy by 3.2% compared with traditional models, enabling more accurate prediction of potential risks in crowd gathering scenarios.