Fengyuan Liu

Phone:+86 1834800001 | Mail: oxfengyuan@gmail.com

Education Background

University of Oxford, Oxford, England 10/2022-10/2023

- M.Sc. Advanced Computer Science
- Topics Covered: Safety of Foundation Model, Multimodal AI

University of Washington, Seattle, WA 09/2017-12/2020

- B.Sc. in Computer Science
- B.Sc. in Applied Computational Mathematical Science (Data Science & Statistics)
- GPA: 3.95/4.0 (around top 1%)
- Topics Covered: Machine Learning, Deep Learning, Stochastic Process, Cryptography

Publications

[1] **Which Model Generated This Image? A Model-Agnostic Approach for Origin Attribution**
Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, Jindong Gu
European Conference on Computer Vision ***(ECCV)***, 2024

[2] **An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models**
Haochen Luo*, Jindong Gu*, **Fengyuan Liu**, Philip Torr
International Conference on Learning Representations ***(ICLR)***, 2024

[3] **DrugGPT: A Knowledge-Grounded Collaborative Large Language Model for Faithful and Evidence-based Drug Analysis**
Fenglin Liu*, Hongjian Zhou*, Wenjun Zhang, Guowei Huang, Lei Clifton, David Eyre, Haochen Luo, **Fengyuan Liu**, Kim Branson, Patrick Schwab, Xian Wu, Yefeng Zheng, Anshul Thakur, and David A. Clifton
Nature Biomedical Engineering ***(Nat. Biomed. Eng)***, 2024 (under review)

[4] **OpenFE: Automated Feature Generation beyond Expert-level Performance**
Tianping Zhang, Zheyu Zhang, Haoyan Luo, **Fengyuan Liu**, Wei Cao, Jian Li
International Conference on Machine Learning ***(ICML)***, 2023

Research Experience

Department of Engineering Science, University of Oxford Oxford, United Kindom
Research Intern, TVG, under the supervision of Dr. Jindong Gu and Prof. Philip Torr 05/2023–12/2023

Project 1: Which Model Generated This Image? [1]

Research Summary: This project presents a model-agnostic method for origin attribution, focusing on determining whether a given image is generated by a specific model when only limited image samples are available and direct access to model parameters is restricted.

- Developed OCC-CLIP, a CLIP-based framework that enabled few-shot one-class classification.
- Utilized the cross-modal representation capabilities of CLIP to compare image embeddings for model attribution, enhanced by adversarial data augmentation (ADA) to improve performance with limited data.
- Demonstrated high accuracy in identifying source models across 8 generative models; Surpassed baseline methods in robustness and resilience, making OCC-CLIP applicable even in real-world commercial systems like DALL-E 3.

Project 2: An Image is worth 1000 lies [2]

Research Summary: This project introduces the Cross-Prompt Attack (CroPA) to enhance adversarial

transferability across prompts in vision-language models, revealing significant vulnerabilities and advancing insights into model robustness.

- Developed the Cross-Prompt Attack (CroPA) algorithm, leveraging learnable prompt perturbations to enhance adversarial transferability across diverse prompts.
- Tested CroPA on popular vision-language models (Flamingo, BLIP-2, InstructBLIP) and tasks like image classification, captioning, and visual question answering, which demonstrated its superior transferability performance compared to baseline methods.
- Concluded that CroPA revealed significant vulnerabilities in vision-language models, highlighting the need for improved model robustness against adversarial attacks across varied prompts.

Institute for Interdisciplinary Information Sciences, Tsinghua University Beijing, China
Research Intern, ADL Group, under the supervision of Prof. Jian Li 05/2022–09/2022

Project 1: Automatic Feature Generation [4]

Research Summary: This study introduces OpenFE, an automated feature generation tool that leverages efficient feature boosting and two-stage pruning methods to achieve expert-level performance in feature engineering, enhancing machine learning models on complex tabular data.

- Developed OpenFE, utilizing a novel FeatureBoost algorithm to evaluate incremental performance of new features without full model retraining, alongside a two-stage pruning algorithm for efficient feature selection.
- Reproduced AutoCross, AutoFeat, SAFE and FCTree methods and compared them with OpenFE.
- Outperformed baseline methods on 10 benchmark datasets and achieved competitive results against human experts in feature generation, with notable success in high-ranking Kaggle competitions, which demonstrating OpenFE's efficacy in enhancing model performance on tabular data.

Project 2: Smart beta based on multi-factor models

- Pre-processed raw factors in the tabular form about all stocks listed on the Shanghai and Shenzhen stock markets from 2017 to present.
- Dealt with factors by filtering stocks, excluding extreme values, filling null values, doing industry neutral, and standardizing.
- Mainly employed Lightgbm to train and compare the prediction results with different labels (pct1, pct2, or pct5) with various factors combination.

Industry Experience

Tencent Shenzhen, China
Research Intern at Tencent AI lab & Robotics X 01/2024-10/2024

Project 1: Cracking the Collective Mind: Adversarial Manipulation in Multi-Agent Systems

Research Summary: This project presents a M-Spoiler framework, demonstrating that a single manipulated agent can disrupt decision consistency across a multi-agent system, similar to a Byzantine Fault in distributed systems.

- Proposed a research question on the safety of multi-agent systems: If one agent is accessible to attackers, can the decision of the entire multi-agent system be manipulated?
- Formulated the research question as a game with incomplete information, and proposed a framework called M-Spoiler (Multi-agent System Spoiler).
- Conducted experiments on different classification tasks, LLMs (Llama2, Vicuna, etc.), and algorithmic backbones (GCG, AutoDAN, etc.) to demonstrate the effectiveness of the proposed framework and provided insights into mitigating such risks.

Project 2: Improving Factuality and Reasoning in Vision Language Models through Multiagent Debate

Research Summary: Explored the weaknesses of multi-agent, multi-modal systems.

Project 3: Self-play for LLM-based Agent

Research Summary: Investigated alignment through the lens of two-agent games, involving iterative interactions between an adversarial and a defensive agent. Strengthened the defensive agent's resilience to malicious attacks by optimizing reinforcement learning strategies and reward functions.

Talks & Activities & Service

Academic Talks (Presenting or Attending)

Which Model Generated This Image? A Model-Agnostic Approach for Origin Attribution
-- *ECCV 2024, online, Milan, Italy* 10/2024

A conversation with Fosun Group Global Partner Mr. Vincent Li
-- *Oxford Said Business School, Oxford, United Kindom* 03/2023

Structural Deep Learning in Financial Asset Pricing by Jianqing Fan
-- *Department of Statistics, Oxford, United Kindom* 10/2022

Oxford Fintech & Legaltech Society

Research Associate 01/2023-04/2023

- Explore the impact modern technology is having on financial institutions legal services, and regulation.

Studies, Experiments, Applications Academy (organized by students from Keble College)

Honor Scholar of Mathematical Studies & Computer Science Department 09/2022-03/2023

- Guided several teenagers to launch research, helped them explore their interests in the area of STEM.

Conference Reviewer

NeurIPS 2023

Skills & Hobbies

Professional Qualification: CFA Exam Level I: Pass

Computer Skills:

Programming: *Python, Java, C#, JavaScript, R, SQL*

Scientific Computing and Engineering: *MATLAB*

Academic Writing and Document Formtting: *LaTeX*

Hobbies: Chinese Kung Fu, Piano, Swimming, Ancient Chinese Philosophy