Guangyuan Hao

(+1) 51 0775 8838 | guangyuan.hao@connect.ust.hk | guangyuanhao.github.io

Education

Hong Kong University of Science and Technology — *Hong Kong*

Master of Philosophy (MPhil) in Artificial Intelligence — *Sept. 2020 - Sept. 2022*

- Collaborated closely with Prof. Hao Wang (Rutgers University).
- Overall GPA: $3.95/4.30$; Major GPA: $4.05/4.30$
- Received a full two-year scholarship.

University of Electronic Science and Technology of China — *Chengdu, China*

Bachelor of Engineering (B.E.) in Information Display and Optoelectronic Technology — *Sept. 2012 - Jul. 2016*

- GPA: $3.97/4.00, 91.3/100$; Ranking: $1^{st}/184$
- GPA for mathematics and theoretical physics courses: $4.00/4.00$

Experience

Scholar at the ML Alignment & Theory Scholars (MATS) Program — *Berkeley, USA*

Jun. 204-Now

- Conduct AI Safety research focused on eliminating backdoor attacks from LLMs in collaboration with the Center for AI Safety, while also attending talks, workshops, and networking events with other members of the Berkeley alignment research community.

Research Assistant at MBZUAI. — *Abu Dhabi*

Apr. 2023-Jun. 2024

- Engaged in extensive research in the fields of causal inference and causal discovery under the Guidance of Prof. Kun Zhang (MBZUAI & CMU) and Prof. Jiji Zhang (CUHK).

Papers

Natural Counterfactuals With Necessary Backtracking — *Hong Kong & Abu Dhabi*

Guang-Yuan Hao*, Jiji Zhang*, Biwei Huang, Hao Wang, Kun Zhang (* indices equal contribution); **Fortieth International Conference on Machine Learning (ICML 2023)** Workshop on Counterfactuals in Minds and Machines (Oral Presentation); Submitted to NeurIPS 2024 — *Dec. 2022 - Sept. 2023*

- Developed a general framework of what we call natural counterfactuals, which are both more flexible and more realistic than the standard framework.
- Proposed a novel optimization framework for generating natural counterfactuals. By combining a naturalness constraint with a principle of minimal change that discourages unnecessary backtracking, we seek to strike the best balance between natural and non-backtracking counterfactuals.
- Presented a detailed method in the general framework and tested it empirically. The empirical results, on both simulated and real data, demonstrate the efficacy of our method in contrast to non-backtracking counterfactuals.

A Conditional Independence Test in the Presence of Discretization — *Abu Dhabi*

Boyang Sun, Yu Yao, **Guang-Yuan Hao**, Yumou Qiu, Kun Zhang; Submitted to NeurIPS 2024 — *Oct. 2023 - Feb. 2024*

- Developed a CI test for ensuring accurate analysis in scenarios where data has been discretized, which are common due to limitations in data collection or measurement techniques, such as in financial analysis and healthcare.
- Our CI test can handle various scenarios including 1). Both variables and are discretized 2). Both variables and are continuous. 3). One of the variables is discretized.

Composite Active Learning: Towards Multi-Domain Active Learning with Theoretical Guarantees — *Hong Kong*

Guang-Yuan Hao, Hengguan Huang, Haotian Wang, Jie Gao, Hao Wang; **Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2024)** — *Dec. 2021 - May 2023*

- Identified the problem of multi-domain active learning. Proposed Composite Active Learning (CAL) as the first general deep AL method for addressing the problem of multi-domain active learning.
- Provided theoretical guarantees that CAL with our budget assignment strategy achieves a lower generalization error bound than existing AL methods.
- Conducted experiments on both synthetic and real-world datasets with detailed ablation studies, showing that CAL significantly improves performance over the state of the art for multi-domain active learning.
- See details at https://github.com/Wang-ML-Lab/multi-domain-active-learning.

Taxonomy-Structured Domain Adaptation

Hong Kong

Tianyi Liu*, Zihao Xu*, Hao He, **Guang-Yuan Hao**, Guang-He Lee, Hao Wang (***** indices equal contribution);

Nov. 2022 - May 2023

Fortieth International Conference on Machine Learning (ICML 2023)

- Identified the problem of domain adaptation (DA) across taxonomy-structured domains and developed taxonomy-structured domain adaptation (TSDA) as the first general DA method to address this problem.
- Built on the classic adversarial framework and introduced a novel taxonomist, which competes with the adversarial discriminator to preserve the taxonomy information.
- Autonomously formulated a core theory to identify conditions under which data features retain taxonomy information.
- See details at https://proceedings.mlr.press/v202/liu23ap/liu23ap.pdf and https://github.com/wang-ML-Lab/TSDA.

Domain-Indexing Variational Bayes: Interpretable Domain Index for Domain Adaptation

Hong Kong

Zihao Xu*****, **Guang-Yuan Hao***, Hao He, Hao Wang (***** indices equal contribution); **Eleventh International Conference on Learning Representations (ICLR 2023)** (Spotlight Presentation)

Dec. 2021 - Sept. 2022

- Identified the problem of inferring domain indices as latent variables in domain adaptation.
- Provided a rigorous definition of "domain index", and developed the first general method, dubbed variational domain indexing (VDI), for inferring such domain indices.
- Provided theoretical guarantees that VDI's final objective function is equivalent to inferring the optimal domain indices.
- Conducted experiments on synthetic and real-world datasets, showing that VDI can infer non-trivial domain indices, significantly improving performance over state-of-the-art domain adaptation methods.
- See details at https://openreview.net/pdf?id=pxStyaf2oJ5 and https://github.com/wang-ML-Lab/VDI.

DyLex: Incorporating Dynamic Lexicons into BERT for Sequence Labeling

Hong Kong

Baojun Wang*, Zhao Zhang*, Kun Xu*, **Guang-Yuan Hao**, Yuyang Zhang, Lifeng Shang, Linlin Li, Xiao Chen, Xin Jiang, Qun Liu (* indicates equal contribution); **2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), Long Paper**

Nov. 2020 - Jan. 2021

- Improved sequence labeling of BERT by combining existing lexicons.
- Proposed a general framework for effectively introducing external lexical knowledge into sequence labeling tasks, and devised a novel knowledge denoising module to fully use large-scale lexicons.
- Our framework outperforms strong baselines and achieves SOTA results on multiple sequence labeling tasks. Our framework has already supported dynamic updates of lexicons to facilitate industrial deployment.
- Achieved the majority of the experiments outlined in this paper and conducted a comprehensive analysis of the experimental results.
- See details at https://github.com/huawei-noah/noah-research/tree/master/NLP/dylex.

MIXGAN: Learning Concepts from Different Domains for Mixture Generation

Guangzhou, China

Guang-Yuan Hao, Hongxing Yu, Weishi Zheng; **27th International Joint Conference on Artificial Intelligence (IJCAI 2018)** (Oral Presentation)

Oct. 2017 - Feb. 2018

- Made an interesting attempt on mixture generation, i.e., absorbing different image concepts (e.g., content and style) from different domains, and thus generating a new domain which the proposed model never observes in the training stage.
- Proposed an unsupervised method to generate new domains by learning and mixing different concepts from different domains. The experimental results demonstrate the effectiveness of MIXGAN, while the related state-of-the-art models fail to generate new domains.
- Other Applications: image-to-image translation and learning a joint distribution of two domains.
- See details at https://www.ijcai.org/proceedings/2018/0306.pdf and https://github.com/GuangyuanHao/MIXGAN.

Honors & Awards

2020-2022	**Postgraduate Studentship**, 220,200 HKD per year, Hong Kong University of Science and Technology
2015	**China National Scholarship**, 1%, 8,000 CNY, the Ministry of Education, China
2015	**Outstanding Graduates**, 1%, University of Electronic Science and Technology of China
2014	**Sekorm First-class Scholarship**, 2%, 8,000 CNY, SEKORM LIMITED
2013	**People's First-class Scholarship**, 10%, 3,000 CNY, University of Electronic Science and Technology of China

Profession & Skills

Peer Reviewer: ICCV 2023, NeurIPS 2023, ICML 2024

Programming and Related: Linux, Python, PyTorch, TensorFlow, LaTeX , MATLAB

Hobbies: Long-Distance Running, Walking, Meditation