

张翼

电话: 18800137987

邮箱: yi.zhang@bjtu.edu.cn

主页: <https://yiiiizhang.github.io>

28 岁, 1996 年 10 月出生



教育经历

2018.9 - 2023.6 北京交通大学

计算机科学与技术 博士 导师: 桑基韬 教授

2014.9 - 2018.6 华北理工大学

计算机科学与技术 本科 Top1% 省优秀毕业生

研究主题

科学问题

计算机视觉以及多模态数据中“虚假相关性”的因果建模和虚假相关性消除, 其中的虚假相关性是人类视角的概念。此科学问题进一步抽象为“如何让深度模型按照人类注入的意图去建模与推理”

落地场景

以“高可靠性的数据挖掘”作为实际问题进行落地研究, 人类注入的意图为“模型应该不使用有偏的规则进行建模”, 包括单模态任务中算法偏见消除以提升模型准确率和鲁棒性, 以及发现多模态预训练模型中对虚假相关性的学习并加以消除

工作/实习经历

华为, 杭州 (算法研究员/博士后)

2023.8 - 至今

大模型研究

鹏城国家实验室, 深圳

2020.9 - 2022.12

预训练大模型可信赖研究

(1) 提升算法鲁棒性: 模型在具体任务上会对不同社会群体表现出不同的决策效果, 提升算法公平性对“预训练信赖”和“增强 OOD 泛化性”都有意义; (2) 探索良性对抗攻击在数据增强和数据编辑上的应用: 发现了对抗攻击在坏的一面“模型的对抗攻击脆弱性”之外, 还有好的一面“良性对抗攻击”, 将对抗攻击应用于数据增强图像信息编辑中; (3) 引导模型按照人类意图进行推理: 基于贝叶斯建模模型意图与人类意图的差异, 并与推断进行纠正; (4) 检测和消除“视觉-语言预训练模型”中学到的虚假相关性, 以社会偏见为应用场景, 检测和消除了预训练模型中的虚假相关性

科研项目

数据反馈和知识融合的跨媒体因果推断, 国家重点研发计划

主要参与人

包括: 探索知识与数据抽象、反馈的协同机制, 建立知识引导的因果推断理论和方法

个人陈述

在近两年时间, 我主要关注大模型领域的研究进展, 包括多模态大模型的预训练与下游任务。其中, 我们将预训练大模型看作 Model as a Service 中的基础模型, 并探索了多模态大模型内部概念与概念之间的关联建模方式, 以及模态与模态之间的差异性等。

此外, 我的博士课题可凝练为“如何让深度模型按照人类注入的意图建模与挖掘数据”这样的科学问题, 旨在通过基于人类知识引导, 让模型尽可能地从不鲁棒分布的数据中挖掘足够多的信息, 而又不会产生可信赖问题。

研究成果

- 讲习班讲者-*Tutorial*

Yi Zhang: Trustworthy Multimedia Analysis. *ACM Multimedia 2021 (CCF-A)*.
<https://2021.acmmm.org/tutorials>

- 论文

Yi Zhang and Jitao Sang. 2020. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. *ACM Multimedia 2020*: 4346-4354 (**CCF-A**)

Yi Zhang, Junyang Wang, and Jitao Sang. 2022. Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models. *ACM Multimedia 2022*: 4996-5004 (**CCF-A**)

Yi Zhang, Jitao Sang et al. 2023. Benign Shortcut for Debiasing: Fair Visual Recognition via Intervention with Shortcut Features. *ACM Multimedia 2023*, 8860-8868(**CCF-A**)

Yi Zhang, Zhefeng wang et al. 2024. Poisoning for Debiasing: Fair Recognition via Eliminating Bias Uncovered in Data Poisoning. *ACM Multimedia 2024*,(**CCF-A**)

Junyang Wang, Ming Yan, **Yi Zhang**, Jitao sang et al. 2023. From Association to Generation: Vision-free-training Captioning Method by Zero-shot Cross-modality Mapping. *IJCAI 2023*: 4326-4334 (**CCF-A**)

Xiaowen Huang, Jiaming Zhang, **Yi Zhang**, Xian Zhao, and Jitao Sang. 2021. Trustworthy Multimedia Analysis. *ACM Multimedia 2021*: 5667-5669 (**CCF-A**)

Shangxi Wu, Qiuyang He, **Yi Zhang**, Dongyuan Lu, Jitao Sang. Debiasing backdoor attack: A benign application of backdoor attack in eliminating data bias. *Information Sciences (SCI-1 Top, CCF-B)*

Yi Zhang, Xinyu Duan et al. 2024. FairRepair API: Bias Mitigation for Black-box Model APIs. *ACM Multimedia 2024*, Under Review (**CCF-A**)

Yi Zhang, Jitao Sang et al. 2024. Inference-Time Rule Eraser: Fair Recognition via Distilling and Removing Biased Rules. *IEEE trans. on Multimedia*, Under Review (**SCI-1 Top, CCF-B**)

- 专利

张翼, 桑基韬 等. 对图像数据进行无偏见分类的方法. 中国. ZL201911099709.0 , 2021 年 10 月 08 日

教学活动

- 助教, Machine Learning, 桑基韬. 北京交通大学, 2020 秋
- 助教, Machine Learning, 桑基韬 & 景丽萍. 北京交通大学, 2019