Yiwei Yang

University of Washington

819 Northeast 67th st
Seattle, WA 98115

yanyiwei.github.io
yanyiwei@uw.edu

Education

10/20 — Present
Seattle, WA
University of Washington
Ph.D. in Information Science
Research Interests: Large Multi-modal Models, Reliability, Spurious Correlations, Hallucinations
Advisor: Bill Howe

09/15 – 05/19
Ann Arbor, MI
University of Michigan
B.S. in Computer Science and Engineering

Professional Experience

10/20 — Present
Seattle, WA
University of Washington
Graduate Student and Researcher

06/22 – 09/22
Remote
SONY AI
Research Intern (Mentors: William Thong, Alice Xiang)

05/18 – 08/18
San Jose, CA
IBM Research, Almaden
Research Intern (Mentors: Eser Kandogan, Prithviraj Sen, Yunyao Li)

Research Projects

UW **Benchmarking Spurious Correlations in Multi-modal LLMs**
Existing spurious correlation benchmarks consider classification setting only. Created a benchmark that surfaces the shortcomings of multi-modal LLMs in a variety of tasks (beyond classification) due to spurious correlations.

UW **Mitigating Spurious Correlations in Multi-modal LLMs**
Introduced a method that enhances the zero-shot capability of multi-modal LLMs with language guidance on spurious attributes, then amplifies the effect of the guidance via contrastive decoding.

UW **Improving Robustness to Spurious Correlations with Concepts**
Introduced a framework that first uses out-of-distribution examples to infer group labels, then applies robust training methods with the inferred group labels to improve worst-group accuracy. Paper
Accepted to CVPR 2024

SONY AI **Bias Propagation in Knowledge Distillation for Vision Models**
Showed that fairness properties transfer from teacher to student model in knowledge distillation.

UW **Surfacing Gender Bias of Vision-Language Models**
Showed that CLIP-based models tend to objectify women through a series of experiments (e.g. embedding association tests).
Preprint - https://arxiv.org/abs/2212.11261
Accepted to FAccT 2023

UW **Surfacing Relations between Distributive and Procedural and Fairness**
Designed a novel fairness loss term using feature attributions for procedural fairness and study how it interacts with distributive fairness.
Accepted to HICSS 2024

Publications

P.11 **Y. Yang.**, A. Liu, R. Wolfe, A. Caliskan, B. Howe. Label-Efficient Group Robustness via Out-of-Distribution Concept Curation. *CVPR 2024*.

P.10 **Y. Yang**, B. Howe. Does a Fair Model Produce Fair Explanations? Relating Distributive and Procedural Fairness. *HICSS 2024*.

P.09 **Y. Yang.**, A. Liu, R. Wolfe, A. Caliskan, B. Howe. Regularizing Model Gradients with Concepts to Improve Robustness to Spurious Correlations. *ICML SCIS 2023*.

P.08 R. Wolfe, **Y. Yang**, B. Howe, A. Caliskan. Contrastive Lanugage-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. *FAccT 2023*.

P.07 **Y. Yang.**, E. Kandogan, Y. Li, W.S.Lasecki, P.Sen. HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop. *ACL 2019*.
(**Best Poster** at Michigan AI Symposium, 1/55)

P.06 **Y. Yang.**, E. Kandogan, Y. Li, W.S.Lasecki, P.Sen. A study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics. *IUI 2019*.

P.05 A. Lundgard, **Y. Yang**, M.L. Foster, W.S. Lasecki. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training.*CHI 2018*.

P.04 S.W. Lee, Y. Zhang, I. Wong, **Y. Yang**, S. D. O'Keefe, W.S. Lasecki. SketchExpress: Remixing Animations for More Effective Crowd-Powered Prototyping Of Interactive Interfaces. *UIST 2017*.

P.03 H. Kaur, M. Gordon, **Y. Yang**, J. Teevan, E. Kamar, J. Bigham, W.S. Lasecki. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. *HCOMP 2017*.

P.02 Y. Chen, S.W. Lee, Y. Xie, **Y. Yang**, W.S. Lasecki, S. Oney. Codeon: On Demand Software Development Assistance. *CHI 2017*.

P.01 S. W. Lee, **Y. Yang**, S. Yan, Y. Zhang, I. Wong, Z. Yan, M. McGruder, C. M. Homan, W. S. Lasecki. Creating Interactive Behaviors in Early Sketch by Recording and Remixing Crowd Demonstrations. *HCOMP 2016*.