Yanghao ZHANG

CONTACT INFORMATION

Tel: +44(0)7422559950 / +86 19849779635

Email: yanghao.zhang@liverpool.ac.uk / yanghao.zhang@outlook.com

Homepage: yanghaozhang.com

Corresponding Address: Flat A1.04, 9 Owen Street, Manchester, M15 4TN, UK

EDUCATIONAL BACKGROUND

PhD Computer Science — *Nov. 2022 - Aug. 2024*
University of Liverpool, UK — Expected

PhD Computer Science — *Sept. 2020 - Nov. 2022*
University of Exeter, UK — Transfer to Liverpool

MSc Artificial Intelligence — *Sept. 2018 - Dec. 2019*
University of Southampton, UK — Distinction

BEng Software Engineering — *Sept. 2014 - June 2018*
Huizhou University, China — GPA: 85%

RESEARCH INTERESTS

Robust Machine Learning, Safety Verification, Computer Vision, Knowledge Graph

SELECTED PUBLICATIONS

X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, X. Zhao, K. Cai, **Y. Zhang**, S. Wu, P. Xu, D. Wu, A. Freitas & M. A. Mustafa. (2024). A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. Artificial Intelligence Review.

T. Zhang, **Y. Zhang**, R. Mu, J. Liu, J. Fieldsend & W. Ruan. PRASS: Probabilistic Risk-averse Robust Learning with Stochastic Search. (IJCAI 2024)

Y. Zhang, T. Zhang, R. Mu, X. Huang & W. Ruan. Towards Fairness-Aware Adversarial Learning. (CVPR 2024)

R. Mu., L. Marcolino, **Y. Zhang**, T. Zhang, X. Huang & W. Ruan. Reward Certification for Policy Smoothed Reinforcement Learning. (AAAI 2024)

T. Zhang, J. Liu, **Y. Zhang**, R. Mu & W. Ruan. DeepGRE: Global Robustness Evaluation of Deep Neural Networks. (ICASSP 2024)

F. Wang, Z. Fu, **Y. Zhang** & W. Ruan. Self-adaptive Adversarial Training for Robust Medical Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention. (MICCAI 2023)

Y. Zhang, W. Ruan, F. Wang & X. Huang. (2023). Generalizing Universal Adversarial Perturbations for Deep Neural Networks. Machine Learning, 112(5), 1597-1626.

F. Wang, **Y. Zhang**, Y. Zheng & W. Ruan. Dynamic Efficient Adversarial Training Guided by Gradient Magnitude. (NeurIPS 2022 Workshop)

Y. Zhang, F. Wang & W. Ruan. Fooling Object Detectors: Adversarial Attacks by Half-Neighbor Masks. (CIKM 2020 Workshop)

Y. Zhang, W. Ruan, F. Wang & X. Huang. Generalizing Universal Adversarial Attacks Beyond Additive Perturbations. (ICDM 2020)

PREPRINT/UNDER REVIEW

Beyond Levels and Continuity: A New Statistical Method for DNNs Robustness Evaluation.

Safeguarding Large Language Models: A Survey. arXiv preprint arXiv:2406.02622.

WORK EXPERIENCE

University of Liverpool *Jan. 2023 - present*
Research Associate (Part-time)

· **FOCETA**: Foundations for Continuous Engineering of Trustworthy Autonomy.
· **EnnCore**: End-to-End Conceptual Guarding of Neural Architectures

Huawei Technologies Research and Development (UK) *May 2021 - Aug. 2021*
Research Intern in Knowledge Graph Team (Part-time)

· Research on editing factual knowledge in large language model without retraining.
· Working on editing knowledge to fix bad cases for semantic parsing.

Huawei Technologies Research and Development (UK) *Nov. 2019 - Jan. 2020*
Software Intern in Knowledge Graph Team (Full-time)

· Deepdive for research survey
· Working on predicting ABox Consistency with Transparent TBoxes via Graph Neural Networks.

Beijing Kunlun Tech Co., Ltd. *Summer 2017*
Internship Engineer (Full-time)

· Developed a crawler programme for data extraction.
· Data analysis and Software testing for mobile game Mabinogi.

TEACHING/RESEARCH ASSISTANT

- Research Assistant, University of Liverpool, 2022-2024

- ECMM422 Machine Learning, University of Exeter, 2021

- ECMM458 Machine Learning (professional), University of Exeter, 2020

- PGR Demonstrator for MSc AI/DS Students, University of Exeter, 2020-2021

ACADEMIC SERVICES

Reviewer for Journal: TKDE/Information Sciences/The Journal of Supercomputing/The Visual Computer

Reviewer for Conference: ECCV/CVPR/ICCV/NeurIPS/CIKM

External Reviewer for Conference: ECML-PKDD/ICML/IJCAI/ICLR

SKILLS

Languages: Native in Mandarin/Teochew; Fluent in English/Cantonese.

Programming Languages: Proficiency in Python, MATLAB. Familiar with PHP, MySQL, HTML/CSS/JavaScript.

Library/Tools: Pytorch, Tensorflow, Scikit-learn, Numpy, Pandas, LaTex, GitHub, Jupyter Notebook, Shell.