张洋豪

联系方式

电话: +44(0)7422559950 / +86 19849779635 (微信同号)

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

主页: yanghaozhang.com

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

教育背景

计算机科学博士 2022 年 11 月 - 2024 年 8 月
利物浦大学, 英国 预计

计算机科学博士 Sept. 2020 年 9 月 - 2022 年 11 月
埃克塞特大学, 英国 转学到利物浦

人工智能理学硕士 2018 年 9 月 - 2019 年 12 月
南安普顿大学, 英国 Distinction

软件工程工学学士 2014 年 9 月 - 2018 年 6 月
惠州学院, 中国 GPA: 85%

研究方向

对抗攻击/防御，安全验证和鲁棒机器学习应用

精选刊物

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)

预印版

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

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

工作经历

利物浦大学 2023 年 1 月 - 至今
研究助理 (兼职)

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

华为爱丁堡研究所 (英国) 2021 年 5 月 - 2021 年 8 月
知识图谱团队的研究实习生（兼职）

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

华为爱丁堡研究所 (英国) 2019 年 11 月 - 2020 年 1 月
知识图谱团队软件实习生 *(全职)*

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

昆仑万维 2017 年暑假
实习工程师 (全职)

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

教学/研究助理

- 研究助理, 利物浦大学, 2022 年 - 2024 年

- ECMM422 Machine Learning 助教, 埃克塞特大学, 2021 年

- ECMM458 Machine Learning (professional) 助教, 埃克塞特大学, 2020 年

- 人工智能/数据科学硕士生的研究生助教, 埃克塞特大学, 2020 年 - 2021 年

学术服务

期刊审稿人: TKDE/Information Sciences/The Journal of Supercomputing/The Visual Computer

会议审稿人: ECCV/CVPR/ICCV/NeurIPS/CIKM

外部会议审稿人: ECML-PKDD/ICML/IJCAI/ICLR

技能

语言: 中文/潮汕话 (母语)，英文/粤语 (熟练).

编程语言: Proficiency in Python, MATLAB. Familiar with PHP, HTML/CSS/JavaScript, MySQL.

工具库: Pytorch, Tensorflow, Scikit-learn, Numpy, Pandas, LaTex, GitHub, Jupyter Notebook, Shell.