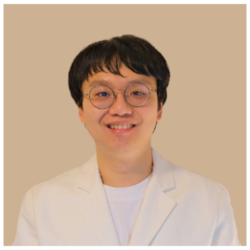
Huan Zhang
Assistant Professor
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
huanz@illinois.edu
https://huan-zhang.com



Professional Preparation

M.S. in Computer Engineering, University of California, Davis	2014
Ph.D. in Computer Science, University of California, Los Angeles	2020

Appointments

Assistant Professor, *Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign (UIUC)* 2023 - present

Postdoctoral fellow, *Department of Computer Science, Carnegie Mellon University (CMU)* 2021 - 2023

Research Areas

Trustworthy Machine Learning; Adversarial Attacks and Defenses for Machine Learning; Computer Security; Formal Methods; Optimization.

Selected Awards

- Schmidt Science AI2050 Early Career Fellow, 2023.
- Winners of the Second, Third, Fourth, and Fifth International Verification of Neural Networks Competition (VNN-COMP), 2021, 2022, 2023, and 2024 (Team lead of the winning tool, α,β-CROWN).
- Adversarial Machine Learning (AdvML) Rising Star Award (sponsored by MIT-IBM Watson AI Lab), 2021
- IBM Ph.D. Fellowship, 2018

Selected Service and Synergistic Activities

- Invited Tutorial Session on *Training and Verification for Learning-based Control* at Modeling, Estimation and Control Conference (MECC) 2024.
- Invited Talk, title "α,β-*CROWN: A Formal Verification Framework for Neural Networks with Applications in Control and Planning*" at INFORMS Annual Meeting, 2024.
- Invited Talk, title "*Solving Large-Scale Non-convex Optimization Problems in Neural Network Verification*", INFORMS Optimization Society Conference, 2024.
- Co-organizer of the 1st and 2nd *Workshop on Formal Verification of Machine Learning*", in conjunction with International Conference on Machine Learning (ICML) 2022, 2023.
- Co-organizer of *Workshop on Socially Responsible Machine Learning*, in conjunction with the International Conference on Learning Representations (ICLR) 2022.
- Co-organizer of the workshop *Trustworthy and Socially Responsible Machine Learning*, in conjunction with the Conference on Neural Information Processing Systems (NeurIPS) 2022.

- Co-organizer of *Workshop on Security and Reliability of Machine Learning*, in conjunction with 19th International Symposium on Automated Technology for Verification and Analysis (ATVA 2021).
- *Tutorial*: "Formal Verification of Deep Neural Networks: Theory and Practice," presented at the 36th AAAI Conference on Artificial Intelligence (AAAI 2022) (tutorial materials available online at https://neural-network-verification.com)
- *Open Source Software:* Since 2021, I have created and been leading the development of α,β-CROWN (https://abcrown.org), an award-winning verification toolbox for rigorously proving the safety of deep neural networks. It won over 20 tools in three consecutive years of International Verification of Neural Networks Competitions.
- Guest Lecture at Yale University, "Formal Verification and Adversarial Attacks of Neural Networks", for course CPSC 680: Trustworthy Machine Learning (2023)
- Guest Lecture at UIUC, "Formal Verification of Deep Neural Networks: Challenges and Recent Advances", for course CS 562: Advanced Topics in Security, Privacy and Machine Learning (2022)
- Guest Journal Editor, *Trustworthy Machine Learning Research Topic*, Frontiers in Big Data, 2021

Relevant Funded Projects

1. National Science Foundation, Safe Learning Enabled Systems (SLES): Verifying and Enforcing Safety Constraints in AI-based Sequential Generation, 2023 - 2026
2. Schmidt Science: AI 2050 Early Career Fellowship (awarded with a $300,000 research grant on AI safety), 2023 - 2025
3. Toyota Research Institute: Model-Based Planning Using Learned AI Models for Robotics, 2024 - 2025

Selected Publications ("*" indicates co-first authors)

1. Regularizing Hidden States Enables Learning Generalizable Reward Model for LLMs. R. Yang, R. Ding, Y. Lin, H. Zhang, T. Zhang. Advances in Neural Information Processing Systems (**NeurIPS**), 2024.
2. D. Zhou, C. Brix, G.A. Hanasusanto, H. Zhang. Scalable Neural Network Verification with Branch-and-bound Inferred Cutting Planes. Advances in Neural Information Processing Systems (**NeurIPS**), 2024.
3. J. Wu, H. Zhang, Y. Vorobeychik. Verified Safe Reinforcement Learning for Neural Network Dynamic Models. Advances in Neural Information Processing Systems (**NeurIPS**), 2024.
4. S. Lin, H. He, T. Wei, K. Xu, H. Zhang, G. Singh, C. Liu, C. Tan. NN4SysBench: Characterizing Neural Network Verification for Computer Systems. Advances in Neural Information Processing Systems (**NeurIPS**), 2024.
5. L. Yang, H. Dai, Z. Shi, C.J. Hsieh, R. Tedrake, H. Zhang. Lyapunov-stable Neural Control for State and Output Feedback: A Novel Formulation for Efficient Synthesis and Verification. International Conference on Machine Learning (**ICML**), 2024.

6. X. Guo, F. Yu, H. Zhang, Lianhui Qin, Bin Hu. COLD-Attack: Jailbreaking LLMs with Stealthiness and Controllability. International Conference on Machine Learning (**ICML**), 2024.
7. A.J. Havens, A. Araujo, H. Zhang, B. Hu. Fine-grained Local Sensitivity Analysis of Standard Dot-Product Self-Attention. International Conference on Machine Learning (**ICML**), 2024.
8. S. Kotha, C. Brix, J.Z. Kolter, K. Dvijotham, H. Zhang. Provably Bounding neural network preimages. Advances in Neural Information Processing Systems (**NeurIPS**), 2023.
9. Robust Mixture-of-Expert Training for Convolutional Neural Networks. Y. Zhang, R. Cai, T. Chen, G. Zhang, H. Zhang, P.Y. Chen, S. Chang, Z. Wang, S. Liu. International Conference on Computer Vision (**ICCV**), 2023.
10. L.-C. Lan, H. Zhang, C.J. Hsieh. Can Agents Run Relay Race with Strangers? Generalization of RL to Out-of-Distribution Trajectories. International Conference on Learning Representations (**ICLR**), 2023.
11. Z. Liu, Z. Guo, Z. Cen, H. Zhang, J. Tan, B. Li, D. Zhao. On the Robustness of Safe Reinforcement Learning under Observational Perturbations. International Conference on Learning Representations (**ICLR**), 2023.
12. J. Zhang, Z. Chen, H. Zhang, C. Xiao and B. Li. DiffSmooth: Certifiably Robust Learning via Diffusion Models and Local Smoothing. In 32nd USENIX Security Symposium (**USENIX Security**), 2023.
13. H. Zhang*, S. Wang*, K. Xu*, L. Li, B. Li, S. Jana, C.J. Hsieh, Z. Kolter. General cutting planes for bound-propagation-based neural network verification. Advances in Neural Information Processing Systems (**NeurIPS**), 2022.
14. L.C. Lan, H. Zhang, T.R. Wu, M.Y. Tsai, I. Wu, C.J. Hsieh. Are AlphaZero-like Agents Robust to Adversarial Perturbations? Advances in Neural Information Processing Systems (**NeurIPS**), 2022.
15. Z. Shi, Y. Wang, H. Zhang, Z. Kolter, C.J. Hsieh. Efficiently Computing Local Lipschitz Constants of Neural Networks via Bound Propagation, Advances in Neural Information Processing Systems (**NeurIPS**), 2022.
16. W. Zhou, F. Liu, H. Zhang, Muhao Chen. δ-SAM: Sharpness-Aware Minimization with Dynamic Reweighting. Findings in Empirical Methods in Natural Language Processing (**EMNLP**), 2022.
17. H. Zhang*, S. Wang*, K. Xu, Y. Wang, S. Jana, C.J. Hsieh, Z. Kolter. A Branch and Bound Framework for Stronger Adversarial Attacks of ReLU Networks. International Conference on Machine Learning (**ICML**), 2022
18. T. Chen*, H. Zhang*, Z. Zhang, S. Chang, S. Liu, P.Y. Chen, Z. Wang. Linearity Grafting: Relaxed Neuron Pruning Helps Certifiable Robustness. International Conference on Machine Learning (**ICML**), 2022.
19. J. Li, H. Zhang, C. Xie. ViP: Unified Certified Detection and Recovery for Patch Attack with Vision Transformers. European Conference on Computer Vision (**ECCV**), 2022.
20. F. Wu, L. Li, H. Zhang, B. Kailkhura, K. Kenthapadi, D. Zhao, B. Li. COPA: Certifying Robust Policies for Offline Reinforcement Learning against Poisoning Attacks. International

Conference on Learning Representations (**ICLR**), 2022.

21. S. Wang*, H. Zhang*, K. Xu*, X. Lin, S. Jana, C.J. Hsieh, Z. Kolter. Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Neural Network Robustness Verification. Advances in Neural Information Processing Systems (**NeurIPS**), 2021.
22. Y. Huang, H. Zhang, Y. Shi, Z. Kolter, A. Anandkumar. Training Certifiably Robust Neural Networks with Efficient Local Lipschitz Bounds. Advances in Neural Information Processing Systems (**NeurIPS**), 2021.
23. L. Rice, A. Bair, H. Zhang, Z. Kolter. Robustness between the worst and average case. Advances in Neural Information Processing Systems (**NeurIPS**), 2021.
24. Z. Shi*, Y. Wang*, H. Zhang, J. Yi, C.J. Hsieh. Fast Certified Robust Training via Better Initialization and Shorter Warmup, Advances in Neural Information Processing Systems (**NeurIPS**), 2021.
25. H. Zhang*, H. Chen*, D. Boning, C.J. Hsieh. Robust Reinforcement Learning on State Observations with Learned Optimal Adversary. International Conference on Learning Representations (**ICLR**), 2021.
26. K. Xu*, H. Zhang*, S. Wang, Y. Wang, S. Jana, X. Lin, C.J. Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. International Conference on Learning Representations (**ICLR**), 2021.
27. C. Zhang, J. Zhao, H. Zhang, K.W. Chang, C.J. Hsieh. Double Perturbation: On the Robustness of Robustness and Counterfactual Bias Evaluation. Annual Conference of the North American Chapter of the Association for Computational Linguistics (**NAACL**), 2021.
28. H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, C.J. Hsieh. Towards Stable and Efficient Training of Verifiably Robust Neural Networks. International Conference on Learning Representations (**ICLR**), 2020.
29. Z. Shi, H. Zhang, K.W. Chang, M. Huang, C.J. Hsieh. Robustness Verification for Transformers. International Conference on Learning Representations (**ICLR**), 2020
30. H. Zhang*, H. Chen*, C. Xiao, B. Li, M. Liu, D. Boning, C.J. Hsieh. Robust Deep Reinforcement Learning Against Adversarial Perturbations on State Observations. Advances in Neural Information Processing Systems (**NeurIPS**), 2020.
31. K. Xu*, Z. Shi*, H. Zhang*, Y. Wang, M. Huang, K.-W. Chang, B. Kailkhura, X. Lin, C.J. Hsieh. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond. Advances in Neural Information Processing Systems (**NeurIPS**), 2020
32. C. Zhang, H. Zhang, C.J. Hsieh. An Efficient Adversarial Attack for Tree Ensembles. Advances in Neural Information Processing Systems (**NeurIPS**), 2020.
33. Y. Wang, H. Zhang, H. Chen, D. Boning and C.J. Hsieh. On ℓ_p-norm Robustness of Ensemble Decision Stumps and Trees. International Conference on Machine Learning (**ICML**), 2020.
34. P.S. Huang*, H. Zhang*, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, P. Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. Empirical Methods in Natural Language Processing (**EMNLP**), 2020.
35. H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning and C.J. Hsieh.

Towards Stable and Efficient Training of Verifiably Robust Neural Networks. International Conference on Learning Representations (**ICLR**), 2020.

36. M. Cheng, J. Yi, P.Y. Chen, H. Zhang and C.J. Hsieh. Seq2sick: Evaluating the Robustness of Sequence-to-sequence Models with Adversarial Examples. AAAI Conference on Artificial Intelligence (**AAAI**), 2020.
37. H. Chen*, H. Zhang*, D. Boning and C.J. Hsieh. Robust Decision Trees Against Adversarial Examples. International Conference on Machine Learning (**ICML**), 2019.
38. H. Zhang, P. Zhang and C.J. Hsieh. RecurJac: An Efficient Recursive Algorithm for Bounding Jacobian Matrix of Neural Networks and Its Applications. AAAI Conference on Artificial Intelligence (**AAAI**), 2019.
39. C.C. Tu, P. Ting, P.Y. Chen, S. Liu, H. Zhang, J. Yi, C.J. Hsieh and S.M. Cheng. Autozoom: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks. AAAI Conference on Artificial Intelligence (**AAAI**), 2019.
40. H. Zhang, H. Chen, Z. Song, D. Boning, I.S. Dhillon and C.J. Hsieh. The Limitations of Adversarial Training and the Blind-spot Attack. International Conference on Learning Representations (**ICLR**), 2019.
41. M. Cheng, T. Le, P.Y. Chen, J. Yi, H. Zhang, and C.J. Hsieh. Query-efficient Hard-label Black-box Attack: An Optimization-based Approach. International Conference on Learning Representations (**ICLR**), 2019.
42. H. Salman, G. Yang, H. Zhang, C.J. Hsieh and P. Zhang. A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks. Advances in Neural Information Processing Systems (**NeurIPS**), 2019.
43. H. Chen*, H. Zhang*, S. Si, Y. Li, D. Boning and C.J. Hsieh. Robustness Verification of Tree-based Models. Advances in Neural Information Processing Systems (**NeurIPS**), 2019.
44. H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck and G. Yang. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. Advances in Neural Information Processing Systems (**NeurIPS**), 2019.
45. J.H. Choi, H. Zhang, J.H. Kim, C.J. Hsieh and J.S. Lee. Evaluating Robustness of Deep Image Super-resolution Against Adversarial Attacks. International Conference on Computer Vision (**ICCV**), 2019.
46. S. Ye, K. Xu, S. Liu, H. Cheng, J.H. Lambrechts, H. Zhang, A. Zhou, K. Ma, Y. Wang and X. Lin. Adversarial Robustness vs. Model Compression, or Both? International Conference on Computer Vision (**ICCV**), 2019.
47. P.Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.J. Hsieh. EAD: Elastic-net Attacks to Deep Neural Networks via Adversarial Examples. AAAI Conference on Artificial Intelligence (**AAAI**), 2018.
48. T.W. Weng*, H. Zhang*, P.Y. Chen, J. Yi, D. Su, Y. Gao, C.J. Hsieh and L. Daniel. Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. International Conference on Learning Representations (**ICLR**), 2018.
49. H. Zhang*, T.W. Weng*, P.Y. Chen, C.J. Hsieh and L. Daniel. Efficient Neural Network Robustness Certification with General Activation Functions. Advances in Neural Information Processing Systems (**NeurIPS**), 2018.

50. D. Su*, H. Zhang*, H. Chen, J. Yi, P.Y. Chen and Y. Gao. Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. European Conference on Computer Vision (**ECCV**), 2018.
51. X. Liu, M. Cheng, H. Zhang, and C.J. Hsieh, 2018. Towards Robust Neural Networks via Random Self-ensemble. European Conference on Computer Vision (**ECCV**), 2018.
52. L. Weng*, H. Zhang*, H. Chen, Z. Song, C.J. Hsieh, L. Daniel, D. Boning and I. Dhillon. Towards Fast Computation of Certified Robustness for ReLU Networks. International Conference on Machine Learning (**ICML**), 2018.
53. H. Chen*, H. Zhang*, P.Y. Chen, J. Yi and C.J. Hsieh. Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning. Annual Meeting of the Association for Computational Linguistics (**ACL**), 2018.
54. P.Y. Chen*, H. Zhang*, Y. Sharma, J. Yi, and C.J. Hsieh. Zoo: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (**AISec**), 2017.

Citation Metrics (November 4, 2024)
Google Scholar: h-index 43, total number of citations 15,000+
A **full list** of publications available at:
https://scholar.google.com/citations?user=LTa3GzEAAAAJ