Shaopeng Fu

Last Update: February 2025

shaopengfu.me
i7cWm4gAAAAJ
github.com/fshp971
shaopeng.fu@kaust.edu.sa
shaopengfu15@gmail.com
+966 (0) 56 534 0337

Research Interests

Foundation Model Security, LLM Jailbreak Attacks, LLM Adversarial Robustness, Deep Learning Theory

Education

King Abdullah University of Science and Technology — **Thuwal, Saudi Arabia**
Ph.D. Student in Computer Science — Aug. 2023 – Present
Advisor: Prof. Di Wang

The University of Sydney — **Sydney, Australia**
Master of Philosophy (Engineering and IT) — Oct. 2019 – Jan. 2021
Advisor: Prof. Dacheng Tao
Thesis: Bayesian Inference Forgetting

South China University of Technology — **Guangzhou, China**
B.Sc in Mathematics and Applied Mathematics — Sep. 2015 – Jun. 2019
Advisor: Prof. Chuhua Xian (Advising the Competitive Programming Group affiliated to School of CSE)
GPA: 3.61/4.00 | Rank: 6/46

Work Experience

JD.com, Inc. — **Beijing, China**
Algorithm Engineer @ JD Explore Academy — Mar. 2021 – Jul. 2022

- First-author of two ICLR 2022 papers on machine learning privacy.
- Co-author of the *White Paper on Trustworthy Artificial Intelligence* with CAICT. (Chn Ver.) (Eng Ver.)
- Chief developer of **TAICore**, a trustworthy AI assessment toolkit powered by JD Explore Academy for assessing the robustness and privacy-preserving ability of white-box and black-box ML models.

Publications

Conferences & Journals

1. **Shaopeng Fu** and Di Wang. **Theoretical Analysis of Robust Overfitting for Wide DNNs: An NTK Approach.** *International Conference on Learning Representation (**ICLR**)*, 2024.
2. **Shaopeng Fu**, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. **Robust Unlearnable Examples: Protecting Data Against Adversarial Learning.** *International Conference on Learning Representation (**ICLR**)*, 2022.
3. **Shaopeng Fu***, Fengxiang He*, and Dacheng Tao. **Knowledge Removal in Sampling-based Bayesian Inference.** *International Conference on Learning Representation (**ICLR**)*, 2022.
4. Zeke Xie, Fengxiang He, **Shaopeng Fu**, Issei Sato, Dacheng Tao, and Masashi Sugiyama. **Artificial Neural Variability for Deep Learning: On Overfitting, Noise Memorization, and Catastrophic Forgetting.** *Neural Computation 33 (8)*, 2021.

Manuscripts

1. **Shaopeng Fu**, Liang Ding, and Di Wang. **"Short-length" Adversarial Training Helps LLMs Defend "Long-length" Jailbreak Attacks: Theoretical and Empirical Evidence.** *arXiv preprint arXiv:2502.04204*, 2025.
2. **Shaopeng Fu**, Xuexue Sun, Ke Qing, Tianhang Zheng, and Di Wang. **Pre-trained Encoder Inference:**

Revealing Upstream Encoders In Downstream Machine Learning Services. *arXiv preprint arXiv:2408.02814*, 2024.

3. Fengxiang He*, **Shaopeng Fu***, Bohan Wang*, and Dacheng Tao. **Robustness, Privacy, and Generalization of Adversarial Training.** *arXiv preprint arXiv:2012.13573*, 2020.

Selected Awards

International Collegiate Programming Contest (ICPC)

- **Silver Medal**, The ICPC Asia-East Continent Final Xi'an Site — Dec. 2018
- **Silver Medal**, The ICPC Asia Regional Contest Qingdao Site — Nov. 2018
- **Gold Medal (Rank: 6/186)**, The ICPC Asia Regional Contest Shenyang Site — Oct. 2018
- **Silver Medal**, The ACM-ICPC Asia Regional Contest Xi'an Site — Oct. 2017

2017-2018 China National Scholarship — Ministry of Education of P.R. China, Nov. 2018

2016-2017 China National Scholarship — Ministry of Education of P.R. China, Nov. 2017

Services

Conference Reviewer

- **ICML** (2022, 2023, 2024, 2025), **ICLR** (2022, 2023, 2024, 2025), **NeurIPS** (2021, 2022, 2023, 2024), **AISTATS** (2021, 2024, 2025).

Conference Committee Member

- **ACM CCS** (2024 Artifact Evaluation), **AAAI** (2025).

Journal Reviewer

- **IEEE TPAMI** (2024), **IEEE TNNLS** (2024), **IEEE TCYB** (2021), **Springer NPL** (2020).

Miscellaneous

Competitive Programming

- Codeforce Profile: https://codeforces.com/profile/fshp971

Programming Languages

- C/C++ (Mainly for Competitive Programming)
- Python (Mainly for AI Research)

Others: PyTorch, JAX, Vim, Docker, Slurm, Linux, Arch Linux