Yangsibo Huang

Curriculum Vitae

Friend Center, 35 Olden Street
Princeton, NJ 08540
☎ *(609) 356-4438*
✉ *yangsibo@princeton.edu*
🖱 *Personal webpage*

Education

2019–Present **Princeton University** *Princeton, NJ*
Ph.D. in Electrical and Computer Engineering
Advisors: Professor Kai Li & Professor Sanjeev Arora
Research interests: Privacy and Security of Machine Learning Systems (e.g., Federated Learning, Large Language Models)

2015–2019 **Zhejiang University** *China*
B.S. in Computer Science & B.A. in Entrepreneurship Management
GPA: 3.95/4, Graduated with Outstanding Honor (Top 1%)

Honors and Awards

2023 Rising Stars in EECS, Year 2023

2023 Wallace Memorial Fellowship *Princeton University*
The highest award conferrable to graduate students in the School of Engineering and Applied Science

2022 The School of Engineering and Applied Science (SEAS) Travel Grant *Princeton University*

2022 The Dean's Fund for Scholarly Travel *Princeton University*

2020 Bell Labs Prize, The Second Place *Bell Labs*
Awarded to our innovations that enhance privacy in distributed deep learning for image and text data

2019 Outstanding Graduate Award *Zhejiang University*
Top 1%

2016 China National Scholarship *Zhejiang University and Chinese government*
Highest scholarship given by Chinese government, top 0.1% nationwide

Experience

05/2023-12/2023 **Google Research** *Mountain View, CA*
Research Intern & Part-time Student Researcher, Hosts: Chiyuan Zhang & Badih Ghazi
Project: Learning with Label Differential Privacy via Projections

10/2022-05/2023 **Google Research** *Remote*
Part-time Student Researcher, Hosts: Chiyuan Zhang & Badih Ghazi
Project: Sparsity-Preserving Differentially Private Training

05/2022-10/2022 **Meta Research** *Bellevue, WA*
Research Intern & Part-time Student Researcher, Host: Seyi Feyisetan
Project: Empirical Privacy Evaluation via Membership Inference and Reconstruction Attacks

09/2018–04/2019 **Harvard Medical School & Massachusetts General Hospital** *Boston, MA*
Visiting Student Researcher, Advisor: Professor Quanzheng Li
Project: Multi-Modality Clinical Data Analysis Using Machine Learning

Conference and Journal Publications

(* means equal contribution, [α] means alphabetical order)

New Preprints

2023 **Detecting Pretraining Data from Large Language Models** *[paper], [code], [web]*
Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, Luke Zettlemoyer

2023 **Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation** *[paper], [code], [web]*
Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, Danqi Chen

2023 [α]**Learning across Data Owners with Joint Differential Privacy** *[paper]*
Yangsibo Huang, Haotian Jiang, Daogao Liu, Mohammad Mahdian, Jieming Mao, Vahab Mirrokni

2023 **kNN-Adapter: Efficient Domain Adaptation for Black-Box Language Models** *[paper]*
Yangsibo Huang, Daogao Liu, Zexuan Zhong, Weijia Shi, Yin Tat Lee

Conference & Journal Publications

2023 **Privacy Implications of Retrieval-Based Language Models** *[paper]*
Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, Danqi Chen
EMNLP 2023

2023 [α]**Sparsity-Preserving Differentially Private Training**
Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chiyuan Zhang
NeurIPS 2023

2022 **Recovering Private Text in Federated Learning of Language Models** *[paper], [code]*
Samyak Gupta*, Yangsibo Huang*, Zexuan Zhong, Tianyu Gao, Kai Li, Danqi Chen
NeurIPS 2022

2021 **Evaluating Gradient Inversion Attacks and Defenses in Federated Learning** *[paper], [code]*
Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, Sanjeev Arora
NeurIPS 2021 (**Oral**, 1% acceptance rate)

2021 **EMA: Auditing Data Removal from Trained Models** *[paper], [code]*
Yangsibo Huang, Xiaoxiao Li, Kai Li
Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021

2020 **TextHide: Tackling Data Privacy in Language Understanding Tasks** *[paper], [code]*
Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, Sanjeev Arora
EMNLP 2020

2020 **InstaHide: Instance-hiding Schemes for Private Distributed Learning** *[paper], [code]*
Yangsibo Huang, Zhao Song, Kai Li, Sanjeev Arora
ICML 2020

2020 **Deepmc: A Deep Learning Method for Efficient Monte Carlo Beamlet Dose Calculation by Predictive Denoising in Magnetic Resonance-Guided Radiotherapy** *[paper]*
Ryan Neph, Qihui Lyu, Yangsibo Huang, You Ming Yang, Ke Sheng
Physics in Medicine & Biology (IF: 3.6, top journal in Medical Physics)

2019 **Deep Q learning Driven CT Pancreas Segmentation with Geometry-aware U-Net** *[paper]*
Yunze Man*, Yangsibo Huang*, Junyi Feng, Xi Li, Fei Wu
IEEE Transactions on Medical Imaging (IF: 10.7, top journal in Medical Image Analysis)

Manuscripts

2020 **Deep Q Deep Learning Based Detection and Localization of Cerebal Aneurysms in Computed Tomography Angiography** *[paper]*

Ziheng Duan, Daniel Montes, Yangsibo Huang, Dufan Wu, Javier M Romero, Ramon Gilberto Gonzalez, Quanzheng Li

2019 **Privacy-Preserving Learning via Deep Net Pruning** *[paper]*

Yangsibo Huang, Yushan Su, Sachin Ravi, Zhao Song, Sanjeev Arora, Kai Li

Talks

11/2023 Advancing Privacy, Safety, and Transparency in Large-Scale Machine Learning Systems
Rice University

11/2023 Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
Princeton Language and Intelligence (PLI) seminar

08/2023 Sparsity-Preserving Differentially Private Training
Privacy-Preserving Machine Learning Workshop 2023

05/2023 Gradient Inversion Attacks in Federated Learning: Generalizing From Image to Text
Zhejiang University

10/2022 Recovering Private Text in Federated Learning of Language Models
Princeton NLP Seminar

06/2022 Gradient Inversion Attacks in Federated Learning: Generalizing From Image to Text
Center for Brain-Inspired Computing, Industry Meeting

05/2022 Gradient Inversion Attacks in Federated Learning: Attacks, Limitations and Defenses
The University of British Columbia

12/2021 Evaluating Gradient Inversion Attacks and Defenses in Federated Learning
NeurIPS 2021, Oral presentation (the Privacy & Fairness track)

11/2020 TextHide: Tackling Data Privacy in Language Understanding Tasks
Princeton NLP Seminar

Teaching and Mentoring

Teaching

- Teaching assistant for ECE 382: Probabilistic Systems and Information Processing (Spring 2021)

Mentoring

- Boyi Wei, PhD Student at Princeton
- Samyak Gupta, PhD Student at Princeton
- Ayush Alag, Undergrad at Princeton → Stanford
- Naomi Boneh, High school student → Stanford
- Emma Hong, High school student → Stanford

Professional Services

Program Committee

- Workshop on Federated Learning for Data Mining, 2023
- Workshop on Federated Learning and Analytics in Practice, 2023
- Workshop on Interpretable Machine Learning in Healthcare, 2021 & 2022
- Workshop on Computer Vision for Automated Medical Diagnosis, 2021

Conference and Journal Reviewer

- ICML, 2021 - 2023
- NeurIPS, 2021 - 2023

- ICCV, 2021 & 2023
- Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2022
- IEEE Transactions on Medical Imaging (TMI), 2021 - 2022

Community Service

- Planning committee for Princeton Graduate Women in Science & Engineering (GWiSE), 2022
- Volunteer for Princeton AI4ALL program for rising 11th graders from underrepresented groups, 2022

Selected Press

2020	Nokia announces 2020 Bell Labs Prize winners	*[Link]*
2020	Bell Lab Prize honors Princeton team for method to meld privacy and deep learning	*[Link]*

Technical Skill

- Programming Language: proficient in Python; basic ability in C, C++, Java and Javascript
- Deep Learning Framework: proficient in PyTorch, Tensorflow