Qian Lou

Senior Research Scientist
Samsung Research AI Center, USA
E-mail: louqian95@gmail.com

EDUCATION

Ph.D.	Computer Engineering	Indiana University Bloomington	2017 - 2021
M.S.	Computer Engineering	Indiana University Bloomington	2017 - 2019
B.S.	Computer Science	Shandong University	2013 – 2017

RESEARCH AREA

- **Privacy-Preserving Machine Learning:** Applied Cryptography, Deep Learning, CV/NLP
- **Machine Learning and Systems:** Model Compression, Algorithm/System Co-Design

HONORS AND AWARDS

2022	Samsung Research America Q4 Best Paper Award
2021	Luddy Outstanding Research Award, Indiana University Bloomington
2020	Young Fellowship, IEEE Design Automation Conference (DAC)
2020	Best Paper Nomination, ACM Parallel Architectures and Compilation Techniques (PACT)
2019	Travel Award, 2019 Conference on Neural Information Processing Systems (NeurIPS)
2018	Best Paper Nomination, 2018 International Conference On Computer Aided Design (ICCAD)

WORKING EXPERIENCE

Samsung Research America Mountain View, CA
Senior Research Scientist June 2021 – Present

Research Goal and Approaches: Developing private, fast, and compact deep learning models for real-world vision and language understanding.

- Compressed the state-of-the-art transformers by 2.1x ~ 8.9x for the deployment on mobile devices.
  - Related publications: [19] [20] [15]
- Developed algorithms for providing deep learning with privacy guarantees.
  - Related publications: [14][23]

Indiana University Bloomington Bloomington, IN
Research Assistant, Intelligent Systems Engineering August 2017 – June 2021

Research Goal and Approaches: Aiming to design efficient and practical privacy-preserving machine learning systems by local computation on mobile systems and secure, confidential computation on servers.

- Developing the first kernel-wise quantization algorithm and applications for deep learning inference on mobile devices.
  - Related publications: [10] [9] [6] [3] [15]

- Developed algorithms and frameworks for providing real-world private deep learning with privacy guarantees, low latencies, and competitive accuracies.
  - Related publications: [23][17][13] [11] [8] [5]
- Developed a framework that significantly improves the practical efficiency of private deep learning by the combination of on-device local computation for nonlinear activation and cloud-based secure computation for linear operations.
  - Related publications: [14] [12] [17]

Samsung Research America Mountain View, CA
Research Intern May 2020 – September 2020

Research Goal and Approaches: Developing a secure, accurate, and fast neural network inference for real-world image classification and speech recognition applications.

- Accelerated on-device neural network inference by only running cheap activation, e.g., *ReLU*, on local devices, and outsourcing privacy-preserving linear layers, e.g., *Convolution*, to the powerful servers. Related publications: [14]

PUBLICATIONS

[23] **Qian Lou**, Bo Feng, Geoffrey C. Fox, and Lei Jiang. " FVNet: A Low-Latency Privacy-Preserving Neural Network ". Under Review.

[22] **Qian Lou**, Yen-Chang Hsu, Burak Uzkent, Ting Hua, Yilin Shen, and Hongxia Jin. " Lite-MDETR: A Lightweight Multi-Modal Detector ". Under Review.

[21] **Qian Lou**, Mengxin Zheng. "Primer: Privacy-Preserving Transformer on Encrypted Data". Under Review.

[20] Yen-Chang Hsu, Ting Hua, Sungen Chang, **Qian Lou**, Yilin Shen, and Hongxia Jin. " Language model compression with weighted low-rank factorization ". In ICLR 2022.

[19] **Qian Lou**, Ting Hua, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. "DictFormer: Tiny Transformer with Shared Dictionary". In ICLR 2022.

[18] Mingqin Han, Yilian Zhu, **Qian Lou**, Zimeng Zhou, Shanding Guo, Lei Ju. "coxHE: A software hardware co-design framework for FPGA acceleration of homomorphic computation". In DATE 2022.

[17] Bo Feng, **Qian Lou**, Geoffrey C. Fox, and Lei Jiang. "Low Latency Privacy-Preserving Text Analysis With GRU". The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[16] **Qian Lou,** Lei Jiang "HEMET: A Homomorphic-Encryption-Friendly Privacy-Preserving Mobile Neural Network Architecture", International Conference on Machine Learning (ICML), 2021

[15] Changsheng Zhao, Ting Hua, Yilin Shen, **Qian Lou**, and Hongxia Jin. "Automatic Mixed-Precision Quantization Search of BERT". In IJCAI 2021.

[14] **Qian Lou**, Yilin Shen, Hongxia Jin, and Lei Jiang. "SAFENet: A Secure, Accurate, and Fast Neural Network Inference". In ICLR 2021.

[13] **Qian Lou**, Wenjie Lu, Cheng Hong, and Lei Jiang. "Falcon: Fast Spectral Inference on Encrypted Data". In NeurIPS 2020.

[12] **Qian Lou**, Bian Song, and Lei Jiang. "AutoPrivacy: Automated Layer-wise Parameter Selection for Secure Neural Network Inference". In NeurIPS 2020.

[11] **Qian Lou**, Bo Feng, Geoffrey C. Fox, and Lei Jiang. "Glyph: Fast and Accurately Training Deep Neural Networks on Encrypted Data". In NeurIPS 2020.

[10] **Qian Lou**, Sarath Janga, and Lei Jiang. "Helix: Algorithm/Architecture Co-design for Accelerating Nanopore Genome Base-calling." In PACT 2020. **Best Paper Nomination (4/197).**

[9] **Qian Lou**, Feng Guo, Minje Kim, Lantao Liu, and Lei Jiang. "AutoQ: Automated Kernel-Wise Neural Network Quantization." In ICLR 2020.

[8] **Qian Lou**, Wenje Lu, Cheng Hong, and Lei Jiang. "HERB: Fast Privacy-Preserving Inference using Block Circulant Weight Matrices". In CCS PPMLP 2020.

[7] Farzinah Zokaee, **Qian Lou**, N. Youngblood, Weichen Liu, and Lei Jiang. "LightBulb: A Photonic-Nonvolatile-Memory-based Accelerator for Binarized Convolutional Neural Networks." In DATE 2020.

[6] **Qian Lou**, Weichen Liu, Wenyang Liu, Feng Guo, and Lei Jiang, "MindReading: An Ultra-Low-Power Photonic Accelerator for EEG-based Human Intention Recognition," In ASP-DAC 2020.

[5] **Qian Lou** and Lei Jiang. "SHE: A Fast and Accurate Deep Neural Network for Encrypted Data." In NeurIPS 2019.

[4] Weichen Liu, Wenyang Liu, Yichen Ye, **Qian Lou**, Yiyuan Xie, and Lei Jiang. "HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers." In DATE 2019.

[3] **Qian Lou**, Wujie Wen, and Lei Jiang. "3DICT: a reliable and QoS capable mobile process-in-memory architecture for lookup-based CNNs in 3D XPoint ReRAMs". In ICCAD 2018.

[2] **Qian Lou**, and Lei Jiang. "BRAWL: A Spintronics-Based Portable Basecalling-in-Memory Architecture for Nanopore Genome Sequencing". IEEE Computer Architecture Letters, 2018.

[1] **Qian Lou**, Mengying Zhao, Lei Ju, Chun Xue, Jingtong Hu, and Zhiping Jia. "Runtime and reconfiguration dual-aware placement for SRAM-NVM hybrid FPGAs." IEEE NVMSA 2017.

TEACHING EXPERIENCE

Associate Instructor Bloomington, IN

ENGR E511: Machine Learning for Signal Processing January 2020 – May 2020

- At Indiana University, I worked as an Associate Instructor (AI) to teach undergraduate and graduate students. I have been an AI for the course *ENGR E511* with about 100 students. Other than grading students' homework, I gained valuable experiences in teaching courses.

Research Mentor of Junior Ph.D. students and internships

Bloomington, IN

Research on Private Deep Learning January 2019 – Present

- I have had wonderful mentoring experiences with junior Ph.D. students. As an example, I will mention my collaboration with Bo Feng. Under my guidance, he was enjoyable to put many efforts on the accuracy verification of private deep learning. We have collaborated on the private deep learning for two years. We published three research papers: NeurIPS 2020, and EMNLP 2021.

PROFESSIONAL SERVICES

- **Conference Reviewing:**
  - FHE.org Organization Committee
  - ACM GLSVLSI TPC member
  - International Conference on Machine Learning (ICML)
  - Conference on Neural Information Processing Systems (NeurIPS)
  - AAAI Conference on Artificial Intelligence (AAAI)
  - ACM Asia and South Pacific Design Automation Conference (ASPDAC)
  - Conference on Computer Vision and Pattern Recognition (CVPR)
- **Journal Reviewing:**
  - ACM Journal on Emerging Technologies in Computing Systems (JETC)
  - IEEE Transactions on Computer Systems

GRANTS APPLICATION EXPERIENCE

NSF CCF-1908992 Bloomington, IN

SHF: Small: Automated Algorithm/Hardware Co-Design for Accelerating Nanopore Base-calling August 2018 – May 2021

- At Indiana University, I worked as a research assistant with Lei Jiang (Principal Investigator) to write proposals and apply for grants. We submitted our proposals to the NSF Software & Hardware Foundation (CHF) foundation at the division of Computing and Communication Foundations (CCF) managed by Sankar Basu and received an award with almost half a million dollars. This awarded research helps us produce multiple publications including [7][8][10][12].

NSF CCF-1909509 Bloomington, IN

FET: Small: A Portable and Intelligent Testing System for Power-efficient and Accurate Foodborne Pathogen Detection Architecture July 2018 – December 2020

- I also made contributions on proposal writing about hardware architecture design for the intelligent testing system. This proposal was submitted to Foundations of Emerging Technologies (FET) managed by Mitra Basu and was awarded with almost half a million dollars. Supported by this grant, two papers [8][11] are published. These grants application experiences foster me to apply independent grants as a Principal Investigator.

REFERENCES

Lei Jiang, Assistant Professor
Department of Intelligent Systems Engineering, Indiana University Bloomington
MESH (2425 N. Milo B Sampson Ln) 114B
(812) 855-7728
jiang60@iu.edu

Geoffrey C. Fox, Distinguished Professor
Department of Intelligent Systems Engineering, Indiana University Bloomington
MESH (2425 N. Milo B Sampson Ln) 158
(812) 856-7977
gcfexchange@gmail.com

Minje Kim, Assistant Professor
Department of Intelligent Systems Engineering, Indiana University Bloomington
Luddy Hall (700 N. Woodlawn Ave) 4140
(812) 856-3675
minje@indiana.edu

Weichen Liu, Nanyang Assistant Professor
School of Computer Science and Engineering, Nanyang Technological University
N4-02b-69b
liu@ntu.edu.sg

Haixu Tang, Professor
Department of Computer Science, Indiana University Bloomington
Lindley Hall 301D
(812) 856-1859
hatang@indiana.edu

Xiaofeng Wang, James H. Rudy Professor
Department of Computer Science, Indiana University Bloomington
Luddy Hall (700 N. Woodlawn Ave) 3046
(812) 856-1862
xw7@indiana.edu