Ran Ran

PhD Student · Computer Science

+1 484-379-2935 | ranranteda@gmail.com | https://github.com/ranran0523 | https://www.linkedin.com/in/ranran0523/

Personal Summary

- **Third year of PhD program**, published **8 papers (4 first-author)**, including 6 top-tier Machine learning, Security, and EDA conference publications-**3 NeurIPS, 1 ICML, 1 AAAI, 1 DAC, 1 ICCAD, and 1 ACSAC**.
- Two-year research experience in privacy-preserving machine learning and solid knowledge of **homomorphic encryption, multi-party computation, differential privacy, and trusted execution environment**.
- Four-year experience and solid knowledge of machine learning and relevant programming frameworks. Strong problem-solving and analytical skills, self-motivated, detail-oriented, and leadership for all projects.
- Strong problem-solving and analytical skills, self-motivated, detail-oriented, and leadership for all projects.
- Coding and fraemworks: Python, C++, Java, MATLAB, SQL, Pytorch, TensorFlow, Keras, Numpy, Matplotlib.

Education

North Carolina State University *Raleigh, NC*
Ph.D. in Computer Science *2023.8 - present*
- Advisor: Dr. Wujie Wen

Lehigh University *Bethlehem, PA*
M.S. in Industrial and System Engineering, Ph.D. in Computer Engineering *2018.8 - 2023.8*
- Advisor: Dr. Martin Takáč, Dr. Wujie Wen

Nankai University *Tianjing, China*
B.S. in Industrial Engineering and Applied Mathematics - Dual Degree *2013.8 - 2018.6*
- Advisor: Dr. Qian Wang

Research Project Experience

Homomorphically Encrypted Inference for GCN-based models *Bethlehem, PA*
Advisor: Dr. Wujie Wen *Jan. 2022 - Nov. 2022*
- The **first** Homomorphic Encryption(HE)-based privacy-preserving machine learning framework for GCN-based models.
- Propose the Adjacency Matrix Aware-data representation format of ciphertexts, with latency speedup up to **3.1**×.
- Further optimize activation layers to trade off smaller cryptographic parameters, with latency speedup up to **2.3**×.
- Develop a parallel-packing format to reduce latency by **5**× when node feature and adjacency matrix encrypted
- **Skill:** Pytorch, Python and C++. This Project leads to 2 publication at **NeurIPS 2022** and **NeurIPS 2023** .

Accelerate Homomorphically Encrypted Inference for CNN models *Bethlehem, PA*
Advisor: Dr. Wujie Wen *June. 2022 - Sept. 2022*
- The **first** data packing and model architecture co-optimizing framework for speedup encrypted inference.
- Co-optimize data encoding format and model architecture to accelerate CNN, with latency speedup up to **10.21**×.
- Further optimize the model sparsity patterns to skip high-latency HE-operation, with latency speedup up to **6.57**×.
- **Skill:** Pytorch, Python and C++. This Project leads to 1 publication at **ICML 2023**.

Reduce Activation Budget to accelerate PPML by MPC/HE *Bethlehem, PA*
Advisor: Dr. Wujie Wen *June. 2022 - May. 2023*
- Design a framework that aims to reduce the overhead of MPC comparison protocols and with latency speedup to **20**×.
- Support the training framework with knowledge distillation to maintain model accuracy after polynomial approximation.
- Design the **first** structurally activation pruning framework for accelerating HE-based inference by **14.2**×.
- **Skill:** Pytorch, Python and C++. This Project leads to 3 publication at **DAC 2023, NeurIPS 2023 and AAAI 2023 workshop**.

Weight-sharing CNNs training with Convolution Kernel Patterns search by RL *Bethlehem, PA*
Advisor: Dr. Wujie Wen *Nov. 2021 - May. 2022*
- Design a framework-EVE to search and train weight-shared CNNs to achieve a tradeoff between accuracy and inference latency.
- Develop an AutoML algorithm to search optimal kernel patterns by RL from a search space up to **85184** pattern combination.
- Optima CNN models generated by EVE is on average **2.5**× faster than the baseline models without pruning and shared weights.
- **Skill:** Pytorch, Tensorflow and Python. This Project leads to 1 publication at **ICCAD 2022**.

Publication List

Ran R, Xu N, Wang W, Quan G, Yin J, Wen W. CryptoGCN: Fast and Scalable Homomorphically Encrypted Graph Convolutional Network Inference. Proc. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). (Acceptance rate: 2665/10411=25.6%).

Ran R, Luo X, Wang W, Liu T, Quan G, Wen W. SpENCNN: Orchestrating Encoding and Sparsity for Fast Homomorphically Encrypted Neural Network Inference. Proc. 40th International Conference on Machine Learning (ICML 2023). (Acceptance rate=27.9%)

Ran R, Liu T, Wang W, Quan G, Wen W. Penguin: Parallel-Packed Homomorphic Encryption for Fast Graph Convolutional Network Inference. Proc. 37th Conference on Neural Information Processing Systems (NeurIPS 2023). (Acceptance rate; 3221/12343=26.1%)

Ran R*, Hongwu P*, Yukui Luo, Jiahui Z, Kiran T, Tong G, Chenghong W, Xiaolin X, Wujie W, Caiwen Ding. LinGCN: Structural Linearized Graph Convolutional Network for Homomorphically Encrypted Inference. Proc. 37th Conference on Neural Information Processing Systems (NeurIPS 2023). (Acceptance rate; 3221/12343=26.1%)

Xu N, Wang B, **Ran R**, Wen W, Venkitasubramaniam P. NeuGuard: Lightweight Neuron-Guided Defense against Membership Inference Attacks. Proc. ACM 38th Annual Computer Security Application Conference (ACSAC 2022), Jun. 2022, pp. 1-14. (Acceptance rate: 73/303=23%).

Islam S, Zhou S, **Ran R**, Jin Y, Wen W, Ding C, Xie M. EVE: Environmental Adaptive Neural Network Models for Low-power Energy Harvesting System. Proc. ACM/IEEE 41st International Conference on Computer-Aided Design (ICCAD 2022), May. 2022, pp. 1-9. (Acceptance rate: 132/586=22.5%).

Hongwu P, Shanglin Z, Yukui L, Nuo X, Shijin D, **Ran Ran**, Jiahui Z, Shaoyi H, Xi X, Chenghong W, Tong G, Wujie W, Xiaolin X, Caiwen D. RRNet: Towards ReLU-Reduced Neural Network for Two-party Computation Based Private Inference. AAAI'2023 Workshop on DL-Hardware Co-Design for AI Acceleration

Hongwu Peng, Shanglin Zhou, Yukui Luo, Nuo Xu, Shijin Duan, **Ran Ran**, Jiahui Zhao, Chenghong Wang, Tong Geng, Wujie Wen, Xiaolin Xu, Caiwen Ding. PASNet: Polynomial Architecture Search Framework for Two-party Computation-based Secure Neural Network Deployment. Proc. 60th ACM/IEEE Design Automation Conference (DAC 2023). (Acceptance rate:263/1156=22.7%)

Maximillian Machado, **Ran Ran**, Liang Cheng. Embedded Crowdsensing for Pavement Monitoring and its Incentive Mechanisms. Machine Learning under Resource Constraints - Applications, pp. 286-297

Professional Service

Conference Review

2023 **ICLR 2024**, Reviewer
2023 **NeurIPS 2023**, Reviewer

Peer-reviewed Journal Review

Journal reviewer for **Neurocomputing**
Journal reviewer for **IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)**
Journal reviewer for **IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)**

Awards, Fellowships, & Grants

2023 **Travel Grant**, North Carolina State University CSC Department *$ 1,000*