Yue (Julien) Niu

Research Assistant, Ph.D. candidate

✉ yueniu@usc.edu | ⌂ https://yuehniu.github.io/homepage/ | yuehniu | yue-niu-a3084216a
| @YuehNiu | Los Angeles, CA, US

Education

University of Southern California (USC) *Los Angeles, US*

PhD candidate in Computer Engineering 2018 - Present

Supervisor: Salman Avestimehr

Northwestern Polytechnical University (NPU) *Xi'an, China*

MS in Electrical Engineering 2015 - 2018

Supervisor: Wei Zhou (NPU), Zhenyu Liu (Tsinghua Univ.), Xiangyang Ji (Tsinghua Univ.)

Northwestern Polytechnical University (NPU) *Xi'an, China*

BS in Electronics 2011 - 2015

Thesis supervisor: Wei Zhou (NPU)

Research Focus

Efficient Privacy-Preserving Machine Learning

- Differentially private model training and inference in heterogeneous platforms with Trusted Execution Environments (TEEs) and GPUs;
- Goal: Design a privacy-preserving training and inference algorithm with *strong privacy protection* (with differential privacy) and *high computing performance*.

Federated Learning at the Edge

- Federated learning of large models at resource-constrained clients;
- Goal: Design a resource-efficient training algorithm that distributes training workloads to multiple clients. Each client only trains a small sub-model.

High-order Stochastic Optimization

- Quasi-Newton methods for large-scale neural network optimization;
- Goal: Design a lightweight and fast quasi-Newton method for distributed neural network optimization at scale.

Neural Network Acceleration

- Accelerate neural network inference on the FPGA platform;
- Goal: Optimize neural network computation and design hardware architecture to accelerate neural network inference.

Experience

Amazon Alexa AI *Los Angeles, CA*

Applied Scientist Intern 06/2022 - 09/2022

Topic: I am working with Furqan Khan and Pradeep Natarajan to design a performance estimation (PE) model to estimate a CV model's performance in the wild. The PE can accurately detect if the CV model gave a correct prediction without resorting to human labeling. **We have paper available at** Here

Amazon Alexa AI *Seattle, WA*

Applied Scientist Intern 06/2021 - 09/2021

Topic: I am advised by Furqan Khan and Pradeep Natarajan to develop efficient object detection DNN models for resource-constrained devices. We managed to use knowledge distillation (KD) to reduce model size while still preserving good detection performance.

Tsinghua University *Beijing, China*

Research Intern 06/2017 - 06/2018

Topic: I am advised by Professor Zhenyu Liu and Xiangyang Ji to design an efficient convolutional neural network (CNN) accelerator on FPGA. We accelerate neural network training from both algorithmic and hardware optimization. Algorithmically, we exploit the low-rank structure in CNNs to reduce computational footprints. For hardware optimization, we design a high-performance convolution unit to over computation and memory access. **A demo is available at** Here

Publications

[1] Sara Babakniya, Souvik Kundu, Saurav Prakash, **Yue Niu**, Salman Avestimehr, *Revisiting Sparsity Hunting in Federated Learning: Why the Sparsity Consensus Matters?*, Transaction on Machine Learning Research (TMLR), 2023. [Link]

[2] **Yue Niu**, Zalan Fabian, Sunwoo Lee, Mahdi Soltanolkotabi, Salman Avestimehr, *mL-BFGS: A Momentum-based L-BFGS for Distributed Large-scale Neural Network Optimization*, Transaction on Machine Learning Research (TMLR), 2023. [Link]

[3] Xiruo Liu, **Yue Niu**, Furqan Khan and Prateek Singhal, *Performance and Failure Cause Estimation for Machine Learning Systems in the Wild*, International Conference on Computer Vision Systems (ICVS), 2023. [Link]

[4] **Yue Niu**, Saurav Prakash, Souvik Kundu, Sunwoo Lee, Salman Avestimehr. *Federated Learning of Large Models at the Edge via Principal Sub-Model Training*, *FL-NeurIPS*, 2022. [Link]

[5] Sara Babakniya, Souvik Kundu, Saurav Prakash, **Yue Niu**, Salman Avestimehr. *Federated sparse training: Lottery aware model compression for resource-constrained edge*, *FL-NeurIPS*, 2022. [Link]

[6] **Yue Niu**, Ramy E. Ali, Salman Avestimehr. *3LegRace: Privacy-Preserving DNN Training over TEEs and GPUs*, *Privacy Enhancing Technologies Symposium (PETs)*, 2022. [Link]

[7] **Yue Niu**, Salman Avestimehr. *AsymmetricML: An Asymmetric Decomposition Framework for Privacy-Preserving DNN Training and Inference*, *ICLR Workshop on Distributed and Private Machine Learning*, 2021. [Link]

[8] **Yue Niu**, Zalan Fabian, Sunwoo Lee, Mahdi Soltanolkotabi, Salman Avestimehr. *SLIM-QN: A Stochastic, Light, Momentumized Quasi-Newton Optimizer for Deep Networks*, *ICML Workshop on the Optimization*, 2021. [Link]

[9] **Yue Niu**, Rajgopal Kannan, Ajitesh Srivastava, Viktor Prasanna. *Reuse Kernels or Activations? A Flexible Dataflow for Low-latency Spectral CNN Acceleration*, *ACM/SIGDA International Conference on Field-Programmable Gate Arrays (FPGA)*(**Oral**), 2020. [Link]

[10] **Yue Niu**, Hanqing Zeng, Ajitesh Srivastava, Kartik Lakhotia, Rajgopal Kannan, Yanzhi Wang, Viktor Prasanna. *SPEC2: SPECtral SParsE CNN Accelerator on FPGAs*, *IEEE International Conference on High Performance Computing (HiPC)*(**Oral**), 2020. [Link]

[11] Wei Zhou, **Yue Niu**, Guanwen Zhang. *Sensitivity-oriented layer-wise acceleration and compression for convolutional neural network*, *IEEE Access*, 2019. [Link]

[12] Chunsheng Mei, Zhenyu Liu, **Yue Niu**, Xiangyang Ji, Wei Zhou, Dongsheng Wang. *A 200MHZ 202.4GFLOPS@10.8W VGG16 Accelerator in XILINX VX690T*, *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*(**Oral**), 2017. [Link]

[13] **Yue Niu**, Chunsheng Mei, Zhenyu Liu, Xiangyang Ji, Wei Zhou, Dongsheng Wang. *Sensitivity-Based Acceleration and Compression Algorithm for Convolutional Neural Network*, *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*(**Oral**), 2017. [Link]

[14] **Yue Niu**, Wei Zhou, Xiaocong Lian, Xin Zhou, Jiamin Yang. *A Stepped-RAM Reading and Multiplierless VLSI Architecture for Intra Prediction in HEVC*, *The Pacific-Rim Conference on Multimedia (PCM)*, 2016. [Link]

Volunteer Services

Programe Committee Member in Academic Conference 2020 - Present

- SIAM International Conference on Data Mining (SDM): 2024

Peer Reviewer in Academic Conferences/Journals 2020 - Present

- IEEE Transactions on Mobile Computing (TMC): 2023 (1 paper)
- International Conference on Learning Representations (ICLR): 2021 (2 papers), 2022 (4 papers)
- Conference and Workshop on Neural Information Processing Systems (NeurIPS): 2023 (6 papers), 2022 (4 papers)
- International Conference on Machine Learning (ICML): 2023 (4 papers)
- Knowledge Discovery and Data Mining (KDD): 2023 (3 papers)

Presentation / Attendance in International Academic Conferences Oct. 2020 - Present

- Poster preesntation at UC Berkeley Simons Institute for the Theory of Computing, May 2023
- Poster presentation at NeurIPS, New Orleans, LA, Nov. 2022
- Oral Presentation at PETs, Sydney, Australia, July 2022
- Poster Presentation at ICLR, Virtual, May 2021

Awards and Honors

Best Poster Award at *USC-Amazon Annual Symposium on Secure and Trusted ML* Los Angeles *April 2023*

Technical Skills

Programming	C, Python, Verilog, ...
Professional Softwares	PyTorch, Tensorflow, Vivado, Linux, ...
Languages	Chinese(Native), English(Fluent)