Yue Niu

Ph.D. candidate at USC

☎ 213-204-0799 | ✉ yueniu@usc.edu | 🏠 https://yuehniu.github.io/homepage | 📍 Los Angeles, CA, US

Education

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

PhD candidate in Computer Engineering 2018 - Present

Focus: Efficient & Private Machine Learning. **Supervisor**: Salman Avestimehr

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

MS in Electrical Engineering 2015 - 2018

Focus: DNN Acceleration. **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 Experience

Efficient Private Machine Learning

- Differentially Private machine learning with improved model utility [3, 4, 11];
- Private machine learning empowered by trusted execution environments (TEEs) [3, 11].

Large Language Models

- Privacy, bias, and fairness in language models [2];
- Fast training and inference via low-rank self-attention [5].

CNN/Transformer Acceleration

- Accelerate sparse neural networks with dedicated hardware [13, 12].
- Fast training and inference via low-rank models and activations [3, 5, 11, 15];
- Memory-efficient training and inference via low-rank/sparse compression [7, 13, 15];

Federated Learning at the Edge

- Federated learning of large models at resource-constrained clients [6, 1, 10, 7];
- Communication-efficient federated learning with sparse training on clients [7].

Efficient High-order Stochastic Optimization

- Distributed large-scale model training with quasi-newton optimization (e.g., ResNet50, Transformers) [8].

Experience

Amazon Alexa AI *Los Angeles, CA*

Applied Scientist Intern: **Performance Monitoring**, **Privacy** 06/2022 - 09/2022

Topic: 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. **Publication available at** ICVS'23

Amazon Alexa AI *Seattle, WA*

Applied Scientist Intern: **Model Compression**, **Knowledge Distillation** 06/2021 - 09/2021

Topic: 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: **DNN Ccceleration**, **Low-Rank Compression** 06/2017 - 06/2018

Topic: Design efficient convolutional neural network (CNN) accelerator. 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** Here

Selected Publications

[1] Sunwoo Lee, Tuo Zhang, Saurav Prakash, **Yue Niu**, Salman Avestimehr, Embracing Federated Learning: Enabling Weak Client Participation via Partial Model Training, IEEE Transactions on Mobile Computing, 2024.

[2] Lei Gao*, **Yue Niu***, Tingting Tang, Salman Avestimehr, Murali Annavaram, Ethos: Rectifying Language Models in Orthogonal Parameter Space, North American Chapter of the Association for Computational Linguistics (NAACL) 2024 | AAAI workshop in Responsible Language Models, 2024 (Spotlight).

[3] **Yue Niu**, Ramy Ali, Saurav Prakash, Salman Avestimehr, All Rivers Run to the Sea: Private Learning with Asymmetric Flows, IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2024.

[4] **Yue Niu***, Tingting Tang*, Salman Avestimehr, Murali Annavaram, Edge Private Graph Neural Networks with Singular Value Perturbation, *Privacy Enhancing Technologies Symposium (PETs)*, 2024.

[5] **Yue Niu**, Saurav Prakash, Salman Avestimehr, ATP: Enabling Fast LLM Serving via Attention on Top Principal Keys, ACL, 2024, Under Review.

[6] **Yue Niu**, Saurav Prakash, Souvik Kundu, Sunwoo Lee, Salman Avestimehr, Overcoming Resource Constraints in Federated Learning: Large Models Can Be Trained with only Weak Clients, Transaction on Machine Learning Research (TMLR), 2023. [Link]

[7] 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]

[8] **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]

[9] 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]

[10] **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]

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

[12] **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]

[13] **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]

[14] 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]

[15] **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]

Volunteer Services

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): 2024 (6 papers), 2023 (4 papers)
- Knowledge Discovery and Data Mining (KDD): 2023 (3 papers)
- SIAM International Conference on Data Mining (SDM): 2024 (3 papers)

Mentorship 2023

- USC Viterbi Graduate Mentor

Selected Talks

Presentation in International Academic Conferences Oct. 2020 - Present

- Poster preesntation at Theory and Applications Workshop (ITA), Feb 2024
- 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, C++, Python, Verilog
Professional Softwares	PyTorch, Tensorflow, Linux, Docker