Yue **Niu**

linkedin.com/in/yue-niu-092ab8176 julienniu.wordpress.com/
yueniu@usc.edu
Los Angeles, CA PhD student in Electrical and Computer Engineering at USC

I am currently a PhD student in Electrical and Computer Engineering in Univ. of Southern California (USC). I received my M.S in 2018 and B.S degree in 2015 in Electrical Engineering, both from Northeastern Polytechnical Univ. in China.

I am working on distributed machine learning, which covers two aspects : privacy-preserving ML in distributed systems; accelerating ML training using various parallelism technique and more efficient optimizer (like quasi-Newton method). One of my current work is to combine hardware (security-enabled CPUs/GPUs) and information theory based algorithm to better assign workloads to each platform (security-enabled CPUs and GPUs). Hence both performance and privacy can be achieved.

Publications

2021 Yue Niu, Salman Avestimehr, AsymmetricML : An Asymmetric Decomposition Framework for Privacy-Preserving DNN Training and Inference, to appear in Distributed and Private Machine Learning (DPML) at ICLR Workshop.

2020 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).

2019 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).

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

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

2016 Wei Zhou, Yue Niu, 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).

Skills

Programmation	**C/C++**, **Python**, Matlab, Verilog
Frameworks	Tensorflow, Pytorch, Caffe
Tools	Git
OS	Mac OS, Ubuntu

Projects

Tsinghua Univ. 2016 - 2017

Convolutional Neural Network (CNN) models are computationally intensive and memory intensive, and are difficult to be deployed on embedded systems for real world applications. In this project, we are exploiting low-rank attributes in convolutional and dense layers, shrinking model parameters by reducing redundancy in kernels. Then we implement our compressed models in FPGA platforms which greatly reduces inference latency without killing accuracy.

CNN Acceleration | Caffe | Verilog | Vivado

Teaching and Mentoring

2019, 2020

Introduction to Digital Circuits, USC, Dept. of Electrical and Computer Engineering

- Digital circuit basics
- Verilog basics
- FPGA development procedure
- Embedded system (PicoBlaze)

FPGA | Verilog | Nexys3