

YUKE WANG

805-259-9421 ◇ yuke_wang@cs.ucsb.edu

EDUCATION

Ph.D. candidate, Computer Science. <i>University of California, Santa Barbara</i>	GPA: 3.86/4.00	09/2018 – Now
B.E. Software Engineering. <i>University of Electronic Science and Technology of China</i>	GPA: 3.93/4.00 , Rank: 10/759	07/2018

SKILLS

Languages: Python, C/C++, CUDA, Linux Shell.
Tools: Pytorch, NVProf, Tensorflow, Scikit-Learn, Vivado HLS, Intel FPGA SDK.

EXPERIENCE HIGHLIGHT

Rich hands-on experience on: 1) designing and optimizing GPU kernels for Graph Neural Networks, such as GCN and GraphSage; 2) optimizing deep neural networks (**DNNs**), such as ResNet and VGG, at the algorithmic level and system level; 3) deploying **machine learning** techniques (random forest, logistic regression, etc.) for **bigdata analysis**.

INTERSHIP

High-Performance Engineer Intern, NVIDIA, US. High-performance computational framework support for genomics applications.	07/2021 – 10/2021
Research Intern, Alibaba, US. Compiler framework and runtime system for Graph Neural Networks (GNNs) acceleration on GPUs.	07/2020 – 10/2020

RESEARCH & PROJECTS

Optimizing CNNs via Sliding-Channel Convolutions 07/2020 – 10/2020
1) Propose and implement the first optimized design for exploring deep separable convolution on CNNs; 2) At the algorithm level, we incorporate a novel sliding-channel convolution (SCC), featured with filter-channel overlapping to balance the accuracy performance and the reduction of computation and memory cost; 3) At the implementation level, we build an optimized GPU-implementation tailored for SCC by leveraging several key techniques, such as the input-centric backward propagation and the channel-cyclic optimization; 4) Integrate the SCC into the existing Pytorch framework as a new type of convolution operator.

Efficient Graph Neural Networks on GPUs 01/2020 – 05/2020
1) Develop a GPU-based GNN acceleration framework. It abstracts and parameterizes GNN execution paradigms (e.g., Sparse Matrix Multiplication, Message Passing), and automatically crafts an optimal execution based on several factors, such as the GNN model architectures and dataset properties; 2) it also incorporates a novel low-cost workload scheduling scheme to exploit the data spatial and temporal locality through graph structural aware thread mapping, which can largely improve the GPU shared memory access efficiency under the large graph and high embedding dimensionality settings.

Dynamic Sparsity-aware Neural Network Inference 11/2019 – 01/2020
1) Develop a light-weighted auxiliary “little” module with random projection and weight quantization for probing Neural Network (NN) layerwise output sparsity to facilitate NN inference acceleration; 2) The

proposed "little" module inference scheme can be easily applied to various types of neural networks, such as CNN, and LSTM. (e.g., on ResNet-18, our proposed method outperforms the state-of-the-art solutions with much higher FLOPs reduction, memory saving and model accuracy).

Performance and Power Efficient Accelerator for GNNs

09/2019 – 11/2019

Improve the aggregation step of GNNs accelerator through on-chip computation reuse; Implement an efficient sparse matrix transformation based on Locality-Sensitive Hashing and Row-Column Reordering to guide PE workload mapping.

Accelerating Distance-related Algorithms on Reconfigurable Hardware

04/2019 – 09/2019

1) Build a compiler framework for accelerating distance-related algorithms (KNN, N-Body Simulation, etc.) on CPU-FPGA platform; 2) Develop a Domain-specific Language (DDSL) and algorithm-hardware co-design to simplify design workflow.

Optimizing CNN Architectures Based on Information Field

02/2019 - 05/2019

Propose a new static analysis method for CNN structure- and kernel- level optimization; The proposed optimization can save 50% FLOPs and 90% parameters on most state-of-the-art CNNs.

FPGA Design Framework for K-means

01/2019 – 04/2019

Build a comprehensive FPGA design framework for K-means acceleration; The proposed design framework combines an algorithmic and a hardware design optimization.

KPynq: A Work-Efficient Triangle-Inequality based K-means on FPGA

10/2018 – 01/2019

Develop a novel design for K-means acceleration on Xilinx Pynq-Z1 FPGA board; KPynq achieves speedup (up to 4.2 \times) and energy-efficiency (up to 218 \times) compared with the CPU baseline.

PUBLICATIONS

[OSDI'21] Yuke Wang, Boyuan Feng, Gushu Li, Shuangchen Li, Lei Deng, Yuan Xie, Yufei Ding. **GNNAdvisor: An Efficient Runtime System for GNN Acceleration on GPUs**. *USENIX Symposium on Operating Systems Design and Implementation*.

[USENIX ATC'21] Boyuan Feng, Yuke Wang, Gushu Li, Yuan Xie, Yufei Ding. **Palleon: A Runtime System for Efficient Video Processing toward Dynamic Class Skew**. *USENIX Annual Technical Conference*.

[CCGrid'21] Yuke Wang, Boyuan Feng, Gushu Li, Georgios Tzimpragos, Lei Deng, Yuan Xie, Yufei Ding. **TiAcc: Triangle-inequality based Hardware Accelerator for K-means on FPGAs**. *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*.

[ICASSP'21] Boyuan Feng, Yuke Wang, Yufei Ding. **Sparse Adversarial Attack on EEG-based Brain Computer Interface**. *IEEE International Conference on Acoustics, Speech, Signal Processing*.

[AAAI'21] Boyuan Feng, Yuke Wang, Yufei Ding. **UAG: Uncertainty-aware Attention Graph Neural Network for Defending Adversarial Attacks**. *AAAI Conference on Artificial Intelligence*.

[PPoPP'21] Boyuan Feng, Yuke Wang, Guoyang Chen, Weifeng Zhang, Yuan Xie, Yufei Ding. **TCVM: Accelerating Scientific Computing on Tensor Cores with Extended Precision**. *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*.

[IPDPS'21] Yuke Wang, Boyuan Feng, Yufei Ding. **DSXplore: Optimizing Convolutional Neural Networks via Sliding-Channel Convolutions**. *IEEE International Parallel Distributed Processing Symposium*.

[ICTAI'20] *Boyuan Feng, *Yuke Wang, Xu Li, Shu Yang, Xueqiao Peng, Yufei Ding. **SGQuant: Squeezing the Last Bit on Graph Neural Networks with Specialized Quantization.** *International Conference on Tools with Artificial Intelligence*. *: equal contribution.

[ICML'20] Liu Liu, Lei Deng, Zhaodong Chen, Yuke Wang, Shuangchen Li, Jingwei Zhang, Yihua Yang, Zhenyu Gu, Yufei Ding, Yuan Xie. **Boosting Deep Neural Network Efficiency with Dual-Module Inference.** *International Conference on Machine Learning*.

[TCAD'21] Xiaobing Chen, Yuke Wang, Xinfeng Xie, Xing Hu, Abanti Basak, Ling Liang, Mingyu Yan, Lei Deng, Yufei Ding, Zidong Du, Yunji Chen, Yuan Xie. **Rubik: A Hierarchical Architecture for Efficient Graph Learning.** *IEEE Transactions on Computer Aided Design of Integrated Circuits Systems*.

[FCCM'19] Yuke Wang, Zhaorui Zeng, Boyuan Feng, Lei Deng, Yufei Ding. **KPynq: A Work-Efficient Triangle-Inequality based K-means on FPGA.** *IEEE Symposium on Field-Programmable Custom Computing Machines*.

WORK IN SUBMISSION

[SC'21] Yuke Wang, Boyuan Feng, Yufei Ding. **TC-GNN: Accelerating Sparse Graph Neural Network Computation Via Dense Tensor Core.** *The International Conference for High Performance Computing, Networking, Storage, and Analysis*.

[SC'21] *Yuke Wang, *Boyuan Feng, Yufei Ding. **QGTC: Accelerating Quantized GNN via GPU Tensor Core** *The International Conference for High Performance Computing, Networking, Storage, and Analysis*. *: equal contribution.

[SC'21] *Boyuan Feng, *Yuke Wang, Yufei Ding. **APNN-TC: Accelerating Arbitrary Precision Neural Networks on Ampere GPU Tensor Cores** *The International Conference for High Performance Computing, Networking, Storage, and Analysis*. *: equal contribution.

[CIKM'21] Yuke Wang, Boyuan Feng, Yufei Ding. **Optimizing Convolutional Neural Network Architecture via Information Field.** *International Joint Conference on Artificial Intelligence*.

[TCAD'21] Yuke Wang, Boyuan Feng, Gushu Li, Lei Deng, Yuan Xie, Yufei Ding. **STPAcc: A Compiler-based Framework for Accelerating Distance Algorithms on CPU-FPGA Platforms.** *IEEE Transactions on Computer Aided Design of Integrated Circuits Systems*.

HONORS & AWARDS

2020-2021 Outstanding Publication Award in CS Department of UCSB	06/2021
2020 Summer GSR recipient in CS Department of UCSB	06/2020
2019 Summer GSR recipient in CS Department of UCSB	06/2019
Outstanding Graduates Award of UESTC	10/2017
First-class People's Scholarship (2/20 in the Elite Program)	10/2017
Interdisciplinary Contest In Modeling (ICM) [Honorable Mention]	04/2017
Suzhou Industrial Zone Scholarship (2/20 in the Elite Program)	04/2017
International Software Testing Qualifications Board (Certified Tester) [Foundational Level]	10/2016
First-class People's Scholarship (4/116)	04/2016

EXPERIENCE

Summer Intern at Alibaba US Inc.	Summer. 2020
Teaching Assistant of CS160 (Translation of Programming Languages)	Fall. 2019
Teaching Assistant of CS8 (Python Programming Language)	Summer. 2019
Teaching Assistant of CS16 (C++ Programming Language)	Winter. 2019