Ziyu Ying

W340 Westgate Building, The Pennsylvania State University, University Park, PA 16802
(+1)814-862-8640 | zyying98@gmail.com | linkedin.com/in/zyying | ziyu98.github.io

SHORT BIO

I am a final-year Ph.D. candidate at Pennsylvania State University, focusing on **optimizing the performance and energy efficiency for on-device machine learning and video processing** through **algorithm and system co-design**. I also have hands-on experience in other domains like compiler or HW architecture-driven optimizations and other applications like DLRMs and LLMs. I am seeking SWE/SDE or research roles in the industry.

EDUCATION

The Pennsylvania State University — State College, PA
PhD, Computer Science and Engineering — *Aug. 2019 – Present*
Advisors: Mahmut Taylan Kandemir, Chita R. Das

University of Science and Technology of China (USTC) — Hefei, China
BE, Electronic Engineering and Information Science — *Sep. 2015 – Jun. 2019*
Advisors: Wenyi Zhang, Cong Shen

WORK EXPERIENCE

Google LLC — Mountain View, CA
Software Engineer Intern @ AI Experiences — *May. 2023 – Aug. 2023*

- **Large Language Model (LLM) Profiling:** Investigated and evaluated the large language model inference on edge TPUs using simulators and analytical models, and identified the bottleneck operator(s) which are caused by their memory-bound property.
- **LLM Inference Optimization:** Proposed to offload the computation for the bottleneck layers from TPU to the accelerator (specialized for memory-bound operators).
- **Implementation & Evaluation:** Implemented an automatic tool that takes the LLM and hardware configurations as inputs and returns the optimal computation offloading strategy between edgeTPU and accelerator. Achieved ≈70% end-to-end latency reduction for the tested LLMs.

Meta Platforms, Inc. — Seattle, WA
Software Engineer Intern @ Assistant Platform — *May. 2022 – Aug. 2022*

- **Library Support for Voice Assistant Platform:** Implemented various functionalities for the Assistant framework, such as launching target Apps via Android broadcast or deep link, or resolving entities.
- **Implementation & Verification:** Developed several Apps that can be invoked/enabled by the Assistant framework to test the implemented features.

RESEARCH EXPERIENCE

High Performance Computing Lab, Penn State — State College, PA
Research Assistant — *Aug. 2019 – Present*

- **Algorithm and System Co-Design for Optimizing Emerging Video-based Applications on Edge Devices**
  - ***Efficient Point Cloud (PC) Analysis on Edge Devices***: Speed up the on-device PC-based DNNs inference by improving the structuredness of PC data and utilizing the tensor cores in edge GPUs. Achieved 2.2x speedup and 56% (up to 80%) energy saving.
  - ***Pushing Point Cloud Compression (PCC) to the Edge***: Designed highly parallel compression algorithms to better utilize the edge GPU, along with the approximation technique (for both intra- and inter-PC frame), to improve the performance and energy efficiency of PCC on edge/embedded devices. Achieved 34x speedup and 96% energy saving.
  - ***Optimize DNN Inference on Mobile Devices***: Explored the similarities between 2D video frames using HW codec and motion vectors; enabled energy-efficient DNN inference for 2D videos on handhelds (i.e., Pixel phones) by exploiting the computation reuse opportunities across video frames. Achieved 1.55x speedup and 33% energy saving.
- **Workload Partition and Distribution for DLRM Training on Heterogeneous Systems**
  - ***Cost Estimation for DLRM Training***: Proposed an analytical model to estimate the training cost under various <DLRM components, hardware> mappings on the heterogeneous system; utilized such cost model to guide the workload distribution for DLRM training tasks.

Wireless Information Network Laboratory, USTC Hefei, China

Research Assistant *Jan. 2019 – May 2019*

- **Parameter Estimation under Overestimated Probability Constraints**: Designed robust parameter estimation algorithms to generate more conservative estimation values. Such estimators are essential for applications sensitive to the upper limits of estimated values, such as communication channel capacity estimation.

Wireless@HKU Group, HKU Hong Kong

Research Assistant *Jul. 2018 – Aug. 2018*

- **Data Generation for Edge Training**: Explored how <data transmission latency, training accuracy> are affected by different training data generation approaches: either sending an actual sample from the source or creating a synthetic sample at the edge side.

The Laboratory for Future Networks, USTC Hefei, China

Research Assistant *Sep. 2017 – Dec. 2018*

- **Reinforcement Learning & Wireless Communication**: Investigated multi-armed bandit (MAB) models and their applications in wireless communication.

PUBLICATIONS

[C.7] **Ziyu Ying**, Sandeepa Bhuyan, Yan Kang, Yingtian Zhang, Mahmut T. Kandemir, and Chita R. Das.
EdgePC: Efficient Deep Learning Analytics for Point Clouds on Edge Devices. (**ISCA 2023**).

[C.6] **Ziyu Ying**, Shulin Zhao, Sandeepa Bhuyan, Cyan Subhra Mishra, Mahmut T. Kandemir, and Chita R. Das.
Pushing Point Cloud Compression to the Edge. (**MICRO 2022**).

[C.5] **Ziyu Ying**, Shulin Zhao, Haibo Zhang, Cyan Subhra Mishra, Sandeepa Bhuyan, Mahmut T. Kandemir, Anand Sivasubramaniam, and Chita R. Das.
Exploiting Frame Similarity for Efficient Inference on Edge Devices. (**ICDCS 2022**).

[C.4] Sandeepa Bhuyan, Shulin Zhao, **Ziyu Ying**, Mahmut T. Kandemir, and Chita R. Das.
End-to-end Characterization of Game Streaming Applications on Mobile Platforms. (**SIGMETRICS 2022**).

[C.3] Shulin Zhao, Haibo Zhang, Cyan Subhra Mishra, Sandeepa Bhuyan, **Ziyu Ying**, Mahmut T. Kandemir, Anand Sivasubramaniam, and Chita R. Das.
HoloAR: On-the-fly Optimization of 3D Holographic Processing for Augmented Reality. (**MICRO 2021**).

[C.2] Shulin Zhao, Haibo Zhang, Sandeepa Bhuyan, Cyan Subhra Mishra, **Ziyu Ying**, Mahmut T. Kandemir, Anand Sivasubramaniam, and Chita R. Das.
Déjà view: Spatio-temporal Compute Reuse for Energy-efficient 360° VR Video Streaming. (**ISCA 2020**)

[C.1] Zhiyang Wang, **Ziyu Ying**, and Cong Shen.
OPPORTUNISTIC SPECTRUM ACCESS VIA GOOD ARM IDENTIFICATION. (**GlobalSIP 2018**).

TEACHING EXPERIENCE

The Pennsylvania State University

Teaching Assistant *2020Spring*

CMPSC 200: Programming for Engineers with MATLAB

SERVICE

Conference/Journal Reviewer

IEEE Transactions on Computers 2023, CSAE 2023

AWARDS

ISCA 2023 Student Travel Grant 2023

MICRO 2022 Student Travel Grant 2022

Outstanding Student Scholarship, *University of Science and Technology of China* 2016-2018

SELECTED COURSEWORK

✧Operating System Design ✧Advanced Compiler Construction ✧Introduction to Computer Architecture
✧Algorithm Design and Analysis ✧Large-Scale Machine Learning ✧Computer Vision ✧Image Processing

SKILLS

Languages: C, C++, Python, CUDA, Java, Kotlin, Shell, Matlab

Frameworks & Tools: PyTorch, TensorFlow, TensorFlow Lite, Android SDK, NVIDIA Visual Profiler, Perfetto, Git, PIM Simulator, ffmpeg

Operation Systems: Linux, Windows, MacOS