Yifan Zhao

✉ yifanz16@illinois.edu · https://evzh.net · Evan-Zhao

Education

2019– **University of Illinois at Urbana-Champaign**, *Urbana*, IL
Ph.D. candidate in Computer Science
Advisor: Sasa Misailovic, Vikram Adve

2017–2019 **University of Michigan**, *Ann Arbor*, MI
Bachelor of Science and Engineering in Computer Science

2015–2019 **Shanghai Jiaotong University**, *Shanghai*, China
Bachelor of Science in Electrical and Computer Engineering

Research Focus

Performance Autotuning for Deep Learning Programs

Felix is a tensor program optimization framework that uses *gradient descent* to quickly autotune program schedules. Felix is built on TVM and finds high-performance programs in significantly shorter time than TVM's tuning framework (Ansor). "Felix: Optimizing Tensor Programs with Gradient Descent" is under review at ASPLOS'24.

I am actively extending Felix to incorporate more optimizations, such as better graph-level transformations, multi-GPU scheduling, and accuracy-aware optimizations, and support more tensor compilers such as MLIR and Triton.

Accuracy-aware Optimization of Deep Learning Workloads

ApproxCaliper (MLSys 2023) is an *application-aware neural network optimization framework*. ApproxCaliper targets an application that contains DNN components and captures the application's error tolerance of its DNN outputs, to produce more aggressively pruned (and lower latency) DNNs without hurting application-level output quality.

ApproxTuner (PPoPP 2021) uses autotuning to find configurations that maximize speedup and/or energy savings of an application within a quality threshold. In this work, we propose a novel *predictive autotuning* technique using an accuracy predictor, which brings on average 13× speedup compared to conventional autotuning.

Impact, Contribution

- I am the sole developer of the **Felix** DNN optimization framework, which consists of 13K line of code over the deep learning compiler TVM. Much of the code are dedicated to making Felix generally applicable to a wide range of tensor operators, including user-defined ones.

- Using **ApproxCaliper**, I optimized the software pipeline of a commercial autonomous agriculture robot in collaboration with a robotics startup, Earthsense. Our optimizations allows Earthsense to use more lightweight hardware and reduced the cost of deployed compute hardware by 3×.
- I led the development of the **ApproxHPVM** deep learning compiler in the **HPVM** compiler project, including its two major open-source releases, which is used in research projects at multiple institutes, At IBM Research, ApproxHPVM is now used to compile DNN programs to a custom heterogeneous SoC developed for self-driving vehicles.

Selected Publications

[ASPLOS'24] **Felix: Optimizing Tensor Programs with Gradient Descent.**

Yifan Zhao, Hashim Sharif, Vikram Adve, Sasa Misailovic

[MLSys'23] **ApproxCaliper: Exploiting Application-level Error Resiliency for Optimizing Neural Networks.**

Yifan Zhao*, Hashim Sharif*, Peter Pao-Huang, Vatsin Ninad Shah, Arun Narenthiran, Mateus Valverde Gasparino, Nathan Zhao, Abdulrahman Mahmoud, Sarita Adve, Girish Chowdhary, Sasa Misailovic, Vikram Adve.

[PPoPP'21] **ApproxTuner: a Compiler and Runtime System for Adaptive Approximations.**

Hashim Sharif, **Yifan Zhao**, Maria Kotsifakou, Akash Kothari, Ben Schreiber, Elizabeth Wang, Yasmin Sarita, Nathan Zhao, Keyur Joshi, Vikram Adve, Sasa Misailovic, Sarita Adve.

[VR'23] **Power, Performance, and Image Quality Tradeoffs in Foveated Rendering**

Rahul Singh, Muhammad Huzaifa, Jeffrey Liu, Anjul Patney, Hashim Sharif, **Yifan Zhao**, Sarita Adve

[PLDI'19] **Huron: Hybrid False Sharing Detection and Repair.**

Tanvir Ahmed Khan*, **Yifan Zhao***, Gilles Pokam, Barzan Mozafari, Baris Kasikci.

Talks

- **Felix**: Optimizing Tensor Programs with Gradient Descent
  - Talk to IBM Compiler Team, 12/08/2023, Virtual

- **ApproxCaliper**: Exploiting Application-level Error Resiliency for Optimizing Neural Networks
  - MLSys conference talk, 06/06/2023, Miami Convention Center
  - UIUC Compiler Seminar, 02/07/2022, University of Illinois
  - Illinois Autonomous Farm Workshop, 07/07/2021, University of Illinois
- **ApproxTuner**: a Compiler and Runtime System for Adaptive Approximations
  - Invited talk for CS 598 lecture, 03/10/2022, University of Illinois
  - UIUC Visit Day Poster Session, 03/06/2022, University of Illinois
  - Talk to Qualcomm Compiler Team, 12/13/2020, Virtual
  - Talk to Amazon Compiler Team, 07/24/2020, Virtual

Skills in Languages and Frameworks

Languages

Working languages `Python, C/C++ (w/ CUDA), Rust`;
also notably `Haskell, C#, CMake`.

Frameworks

- Deep learning compilers: TVM, MLIR (IREE), Triton
- General-purpose compilers: LLVM (w/ clang)
- Deep learning frameworks: PyTorch, TensorFlow; familiar with ONNX, cuDNN, oneDNN

Teaching and Mentoring

Teaching Assistant

Fall 2023	UIUC CS 426 Compiler Construction (Sasa Misailovic)
Fall 2018	UMich CS 482 Operating System (Baris Kasikci)

Guest Lectures

Fall 2023	UIUC CS 426 Compiler Construction (2 tutorial lectures)
Spring 2022	UIUC CS 598 Approximate and Probabilistic Computing
Fall 2021	UIUC CS 526 Advanced Compiler Construction

Research Mentoring

2023/08. – present	Univ. of Belgrade Undergraduate student, Vanja Kovinić Collaborating to extend Felix into DNN mixed quantization tuning.

2022/08. – 2023/01 UIUC Ph.D. student, Vimarsh Sathia
Collaborated on an extension of the ApproxCaliper project.

2022/04. – 2023/01 UIUC Undergraduate student, Taksh Pratap Singh
Collaborated on on an accuracy-aware tensor compiler project.

2020/12 – 2022/12 UIUC Undergraduate student, Peter Pao-Huang
Collaborated on the ApproxCaliper project.

2021/12 – 2022/05 UIUC Undergraduate student, Yi Zhou
Collaborated on an accuracy-aware tensor compiler project.

2020/12 – 2022/05 UIUC Undergraduate student, Nathan Zhao
Collaborated on the ApproxTuner and ApproxCaliper projects.

2021/05 – 2021/10 Univ. of Belgrade Undergraduate student, Pavle Divović
Collaborated on Tensor DSL and PyTorch backend in HPVM, and improvement of PyTorch performance portability.

References

Sasa Misailovic
Assistant Professor, Department of Computer Science
University of Illinois at Urbana-Champaign
misailo@illinois.edu

Vikram Adve
Donald B. Gillies Professor, Department of Computer Science
University of Illinois at Urbana-Champaign
vadve@illinois.edu

Sarita Adve
Richard T. Cheng Professor, Department of Computer Science
University of Illinois at Urbana-Champaign
sadve@illinois.edu

Girish Chowdhry
Associate Professor, Agricultural and Biological Engineering and Computer Science
University of Illinois Urbana-Champaign
girishc@illinois.edu