Clemens JS Schäfer

Clemens Jonathan Simon Schäfer · ✉ clemensschaefer1994@gmail.com · clemens-schafer.com

Education

Notre Dame, IN
Jul 2018 – Sept 2023

University of Notre Dame
Doctor of Philosophy Computer Science and Engineering
"Hardware-aware Quantization for Biologically Inspired Machine Learning and Inference"
Doctoral Advisor Dr Siddharth Joshi

London, UK
Sep 2016 – Aug 2017

University College London
Master of Science Computational Finance
Graduated with Distinction

Ingolstadt, DE
Oct 2013 – Aug 2016

Catholic University of Eichstätt-Ingolstadt
Bachelor of Science Business Administration with a Specialisation in Economics
Graduated ranked 2nd

Research Interests

Neural Network Model Quantization · Bio-inspired and Neuromorphic Computing
Hardware Acceleration of Machine Learning Algorithms · Spiking Neural Networks (SNNs)

Experience

San Francisco, CA
Oct 2023 – Present

Googel LLC
Software Engineer, ML Performance Team

- Ideation, design, and implementation of an advanced quantization technique which extracts and processes sparse outliers on SparseCore in parallel to dense low-precision TensorCore operations (writing custom Pallas and SparseCore Mosaic kernels), maximizing hardware utilization.
- Implemented Int8 quantization for Veo3.2 using Pallas; leveraging sub-channel granularity, LoRA and activation quantization fusion to achieve a 6-10% performance gain without quality degradation.
- Design and implementation of a gather kernel for the TPU SparseCore in Mosaic for Mixture-of-Expert models, outperforming the generated XLA code by 6% (compared to XLA when the projected started the gain was 6x).
- Researching and deploying quantization on a fleet-wide scale. Leveraging compiler-based quantization injection for framework agnostic deployment and sensitivity signals to determine layer-wise bit widths for quality neutral quantization.

Notre Dame, IN
Jun 2019 – Sept 2023

University of Notre Dame
Graduate Research Assistant, Department of Computer Science and Engineering

- Researching and developing low-power event-based machine learning algorithms, amongst others for computer vision, by advancing machine learning algorithms with findings from neuroscience (such as neural dynamics).
- Investigating the trade-off between energy and accuracy of efficient quantized (and pruned) neural networks, which can be implemented in digital accelerators with different weight storage methods, resulting in several peer-reviewed papers (publications 4 - 8).
- Analyzing quantization for convolutional spiking neural networks trained with local learning, demonstrating ≈73% memory savings at the cost of ≈1% accuracy, results published and presented at ICONS (see publications).
- Collaborating with researchers from the Department of Electrical Engineering and simulating spiking neural networks with neurons based on properties from FerroFET devices, culminating in a joint journal paper (see publications).

Remote
Aug 2022 – Sept 2023

Google LLC
Part-time Student Researcher, ML Performance Team

- Developed a mixed-precision post-training quantization (PTQ) approach that assigns different numerical precisions to tensors in a network based on their specific needs for a reduced memory footprint and improved latency while preserving model accuracy.
- The method augments estimated Hessians with additional information to capture inter-layer dependencies, enabling fast quantization configuration search (on average six model evaluations).
- Evaluate the method's effectiveness on the ResNet50, MobileNetV2, and BERT models, which demonstrated latency reductions of 25.48%, 21.69%, and 33.28%, respectively while incurring no more than 0.01% accuracy degradation (see publications).

Mountain View, CA
May 2022 – Aug 2022

Google LLC
Research Intern, ML Performance Team

- Researching on automated post-training quantization (PTQ) exploring three sensitivity metrics (quantization error, noise injection, and Hessians) and designing two quantization configuration search algorithms (bisection and greedy).
- Demonstrating 27.59% and 34.31% latency savings on a computer vision and natural language processing task, respectively, using second-order information and a greedy search while guaranteeing no more than 1% accuracy degradation.
- Significant amounts of latency reduction can solely be attributed to the greedy search algorithm, which without any guiding sensitivity information, achieved 26.95% and 32.55% latency reduction (see publications).

Remote
Jun 2021 – Sep 2021

Google LLC
Research Intern, QKeras and Application-Specific Machine Learning Team

- Investigating quantization aware training (QAT) with gradient-based bit width learning given model size constraints.
- Proposing a QAT method with (i) hardware-aware heterogeneous differentiable quantization with tensor-sliced learned precision, (ii) targeted gradient modification for weights and activations to mitigate quantization errors, and (iii) a multi-phase learning schedule to address instability in learning arising from updates to the learned quantizer and model parameters.
- The method established a new Pareto frontier in model accuracy and memory footprint with EfficientNet-Lite0 and MobileNetV2, delivering best-in-class accuracy below 4.3 MB of weights and activations (see publications).

Notre Dame, IN
Sep 2018 – May 2019

University of Notre Dame
Graduate Teaching Assistant, Department of Computer Science and Engineering

- Fall semester for "Introduction to Computing C/C++ Programming" and in the spring semester for "Elements of Computing 2".
- Coordinating undergrad TAs, grading assignments, projects and exams on time, holding office hours to counsel students and lecturing when the instructor was absent.

Relevant Publications

1. Jinuk Kim, Marwa El Halabi, Wonpyo Park, **Clemens JS Schaefer**, Deokjae Lee, Yeonhong Park, Jae W Lee, Hyun Oh Song, "GuidedQuant: Large Language Model Quantization via Exploiting End Loss Guidance", International Conference on Machine Learning, 2025
2. Ibrahim Ahmed, **Clemens JS Schaefer**, Gil Tabak, Denis Vnukov, Zenong Zhang, Felix chern, Anatoliy Yevtushenko, Andy Davis, "EQuARX: Efficient Quantized AllReduce in XLA for Distributed Machine Learning Acceleration", ML for Computer Architecture and Systems (MLArchSys), 2025
3. Martin Schiemer, **Clemens JS Schaefer**, Jayden Parker Vap, Mark James Horeni, Yu Emma Wang, Juan Ye, Siddharth Joshi, "Hadamard Domain Training with Integers for Class Incremental Quantized Learning", Conference on Lifelong Learning Agents (CoLLAs), 2024
4. **Clemens JS Schaefer**, Navid Lambert-Shirzad, Xiaofan Zhang, Chiachen Chou, Tom Jablin, Jian Li, Elfie Guo, Caitlin Stanton, Siddharth Joshi, Yu Emma Wang, "Augmenting Hessians with Inter-Layer Dependencies for Mixed-Precision Post-Training Quantization", arXiv:2306.04879, 2023
5. **Clemens JS Schaefer**, Siddharth Joshi, Shan Li, Raul Blazquez, "Edge Inference with Fully Differentiable Quantized Mixed Precision Neural Networks", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024

6. **Clemens JS Schaefer**, Pooria Taheri, Mark Horeni, Siddharth Joshi, "The Hardware Impact of Quantization and Pruning for Weights in Spiking Neural Networks", IEEE Transactions on Circuits and Systems II: Express Briefs, 2023
7. Weier Wan, Rajkumar Kubendran, **Clemens JS Schaefer**, S. Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, "A compute-in-memory chip based on resistive random-access memory", Nature volume 608, pages504–512 (2022)
8. **Clemens JS Schaefer**, Mark Horeni, Pooria Taheri, Siddharth Joshi, "LSTMs for Keyword Spotting with ReRAM-based Compute-In-Memory Architectures", IEEE International Symposium on Circuits and Systems (ISCAS), 2021
9. **Clemens JS Schaefer**, Siddharth Joshi, "Quantizing Spiking Neural Networks with Integers", International Conference on Neuromorphic Systems (ICONS), 2020
10. **Clemens JS Schaefer**, Patrick Faley, Emre Neftci, Siddharth Joshi, "Memory Organization for Energy Efficient Learning and Inference in Digital Neuromorphic Accelerators", IEEE International Symposium on Circuits and Systems (ISCAS), 2020
11. Sourav Dutta, **Clemens JS Schaefer**, Jorge Tomas Gomez, Siddharth Joshi, Suman Datta, "Supervised Learning in All FeFET-Based Spiking Neural Network: Opportunities and Challenges", Frontiers in Neuroscience, 2020

For more publications and citations see Google Scholar.

Relevant Invited Talks and Conference Presentations

- "The Hardware Impact of Quantization and Pruning for Weights in Spiking Neural Networks", **Clemens JS Schaefer**, Pooria Taheri, Mark Horeni, Siddharth Joshi, IEEE International Symposium on Circuits and Systems (ISCAS), Monterey CA, May 22-24, 2023
- "Augmenting Hessians with Inter-Layer Dependencies for Mixed-Precision Post-Training Quantization", **Clemens JS Schaefer**, Navid Lambert-Shirzad, Xiaofan Zhang, Chiachen Chou, Tom Jablin, Jian Li, Elfie Guo, Caitlin Stanton, Siddharth Joshi, Yu Emma Wang, Google LLC, Mountain View CA, May 2023
- "LSTMs for Keyword Spotting with ReRAM-based Compute-In-Memory Architectures", **Clemens JS Schaefer**, Mark Horeni, Pooria Taheri, Siddharth Joshi, IEEE International Symposium on Circuits and Systems (ISCAS), May 22-28, 2021
- "Quantizing Spiking Neural Networks with Integers" **Clemens JS Schaefer**, SynSense AG, Zurich Switzerland (virtual), Apr 2021
- "Memory Organization for Energy Efficient Learning and Inference in Digital Neuromorphic Accelerators", **Clemens JS Schaefer**, Patrick Faley, Emre Neftci, Siddharth Joshi, IEEE International Symposium on Circuits and Systems (ISCAS), Oct 10-21, 2020
- "Quantizing Spiking Neural Networks with Integers" **Clemens JS Schaefer**, Siddharth Joshi, International Conference on Neuromorphic Systems (ICONS), Jul 28–30, 2020
- "Memory Organization and Structures for On-Chip Learning in Spiking Neural Networks" **Clemens JS Schaefer**, Siddharth Joshi, IEEE 63rd International Midwest Symposium on Circuits & Systems, Aug 9–12, 2020

Relevant Course Work

Hardware Platforms for Deep Learning and Optimization · Advanced Computer Architecture
Machine Learning · Complexity and Algorithms · Operating Systems · VLSI Circuit Design

Skills

Programming Languages: Python esp. JAX and PyTorch (proficient), TensorFlow (proficient), C++ (basic)
Tools: Matlab (intermediate), Cadence Virtuoso (basic), LaTeX(intermediate)
Languages: German (native), English (proficient, TOEFL iBT 112), Spanish (conversational)

Fellowship

May 2023 | **NSF ISCAS 2023 Student Travel Grant**, based on merit of paper
Jul 2020 | **Design Automation Conference (DAC)**, Young Student Fellow
Oct 2013 – Aug 2017 | **Friedrich-Naumann-Foundation for Liberty**, Fellowship Holder
Nov 2014 – Apr 2015 | **German Academic Exchange Service**, Fellowship Holder