

EDUCATION

NOTRE DAME, IN JUL 2018 – SEPT 2023	University of Notre Dame <i>Doctor of Philosophy Computer Science and Engineering</i> “Hardware-aware Quantization for Biologically Inspired Machine Learning and Inference” Doctoral Advisor Dr Siddharth Joshi
LONDON, UK SEP 2016 – AUG 2017	University College London <i>Master of Science Computational Finance</i> Graduated with Distinction
INGOLSTADT, DE OCT 2013 – AUG 2016	Catholic University of Eichstätt-Ingolstadt <i>Bachelor of Science Business Administration</i> with a Specialisation in Economics Graduated ranked 2 nd

RESEARCH INTERESTS

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

EXPERIENCE

MOUNTAIN VIEW, CA OCT 2023 – PRESENT	Google LLC <i>Software Engineer, ML Performance Team</i> <ul style="list-style-type: none">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 <i>Graduate Research Assistant, Department of Computer Science and Engineering</i> <ul style="list-style-type: none">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 $\approx 73\%$ memory savings at the cost of $\approx 1\%$ 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 <i>Part-time Student Researcher, ML Performance Team</i> <ul style="list-style-type: none">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.

FRANKFURT, DE
MAR 2018 – JUL 2018

Ernst & Young GmbH
Quant Team/Data Scientist, Financial Accounting Advisory Services

- Performing benchmark analysis for various financial instruments (e.g. equities, bonds, foreign exchange instruments and futures). Investigating price deviations and anomalies along with visualizing the deviations on a portfolio level and communicating the results with international teams.
- Developing a tool to aid the data handling for equity, bond and forward pricing using technologies such as Python3, Flask and SQLite.
- Carrying out large scale bond price deviation analysis to characterize deviations. Training various machine learning models such as ridge regression, support vector regression and regression trees using python-based libraries, e.g. scikit-learn, NumPy, pandas, matplotlib.

RELEVANT PUBLICATIONS

1. 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
2. **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
3. **Clemens JS Schaefer**, Elfie Guo, Caitlin Stanton, Xiaofan Zhang, Tom Jablin, Navid Lambert-Shirzad, Jian Li, Chiachen Chou, Siddharth Joshi, Yu Emma Wang, “Mixed Precision Post Training Quantization of Neural Networks with Sensitivity Guided Search”, arXiv:2302.01382v2, 2023
4. **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
5. **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

6. 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)
7. **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
8. **Clemens JS Schaefer**, Siddharth Joshi, “Quantizing Spiking Neural Networks with Integers”, International Conference on Neuromorphic Systems (ICONS), 2020
9. **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
10. 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
11. Martin Schiemer, **Clemens JS Schaefer**, Mark James Horeni, Yu Emma Wang, Juan Ye, Siddharth Joshi, “Hadamard Domain Training with Integers for Class Incremental Quantized Learning”, arXiv:2310.03675v1, 2023

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