Maximilian Lam

OBJECTIVE

Scaling machine learning and artificial intelligence through end-to-end system optimization spanning infrastructure, hardware/software co-design, efficient model execution, and the integration of privacy-preserving techniques.

EDUCATION & SCHOLARSHIP

Harvard University Aug 2019 - May 2024
PhD in Computer Science
Thesis: Systems and Algorithms for Efficient, Secure and Private Machine Learning Inference
Thesis Focus: AI Privacy
Committee & Advisors: Professor Michael Mitzenmacher, Professor Edward Suh, Professor Vijay Janapa Reddi, Professor Gu-Yeon Wei, Professor David Brooks

Stanford University Aug 2017 - May 2019
MS in Computer Science
Specialization: Artificial Intelligence

University of California, Berkeley Aug 2013 - May 2017
BA in Computer Science
Graduated with High Distinction

Google Scholar
https://scholar.google.com/citations?user=0tPCcKEAAAAJ&hl=en

PUBLICATIONS, PREPRINTS

Maximilian Lam, Jeff Johnson, Wenjie Xiong, Kiwan Maeng, Udit Gupta, Minsoo Rhu, Hsien-Hsin S Lee, Vijay Janapa Reddi, Gu-Yeon Wei, David Brooks, Edward Suh. *GPU-based Private Information Retrieval for On-Device Machine Learning Inference.* Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2024. https://arxiv.org/abs/2301.10904

Maximilian Lam, Michael Mitzenmacher, Vijay Janapa Reddi, Gu-Yeon Wei, David Brooks. *Tabula: Efficiently Computing Nonlinear Activation Functions for Secure Neural Network Inference.* Transactions on Machine Learning Research (TMLR), 2024. https://arxiv.org/abs/2203.02833

Maximilian Lam, Gu-Yeon Wei, David Brooks, Vijay Janapa Reddi, Michael Mitzenmacher. *Gradient Disaggregation: Breaking Privacy in Federated Learning by Reconstructing the User Participant Matrix.* International Conference on Machine Learning (ICML), 2021 (*long talk*). https://arxiv.org/abs/2106.06089

Maximilian Lam, Zachary Yedidiah, Colby Banbury, Vijay Janapa Reddi. *Quantized Neural Network Inference with Precision Batching.* International Conference on Parallel Architectures and Compilation Techniques (PACT), 2021. https://arxiv.org/abs/2003.00822

Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Keith Achorn, Anjali Gopi, David Kanter, **Maximilian Lam**, Mark Mazumder, Vijay Janapa Reddi *The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage.* NeurIPS 2021 Track Datasets and Benchmarks. https://openreview.net/forum?id=R8CwidgJ0yT

Vijay Janapa Reddi, Brian Plancher, Susan Kennedy, Laurence Moroney, Pete Warden, Anant Agarwal, Colby Banbury, Massimo Banzi, Matthew Bennett, Benjamin Brown, Sharad Chitlangia, Radhika Ghosal, Sarah Grafman, Rupert Jaeger, Srivatsan Krishnan, **Maximilian Lam**, Daniel Leiker, Cara Mann, Mark Mazumder, Dominic Pajak, Dhilan Ramaprasad, J. Evan Smith, Matthew Stewart, Dustin Tingley *Widening Access to Applied Machine Learning with TinyML.* Arxiv Preprint, 2021.

https://arxiv.org/abs/2106.04008

Srivatsan Krishnan*, **Maximilian Lam***, Sharad Chitlangia*, Zishen Wan, Aleksandra Faust, Vijay Janapa Reddi. *QuaRL: Quantization for Fast and Environmentally Sustainable Reinforcement Learning.* Transactions on Machine Learning Research (TMLR), 2022. https://openreview.net/pdf?id=xwWsiFmUEs

Colby R. Banbury, Vijay Janapa Reddi, **Max Lam**, et al. *Benchmarking TinyML systems: Challenges and Direction.* Arxiv Preprint, 2020. https://arxiv.org/abs/2003.04821

Maximilian Lam. *Word2Bits - Quantized Word Vectors.* Arxiv Preprint, 2018. https://arxiv.org/abs/1803.05651

Jian Zhang, **Max Lam**, Stephanie Wang, Paroma Varma, Luigi Nardi, Kunle Olukotun, Christopher Ré. *Exploring the Utility of Developer Exhaust.* Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning (DEEM), 2018. https://dl.acm.org/citation.cfm?id=3209895

Jeffrey Regier, Kiran Pamnany, Keno Fischer, Andreas Noack, **Maximilian Lam**, Jarrett Revels, Steve Howard, Ryan Giordano, David Schlegel, Jon McAuliffe, Rollin Thomas and Prabhat. *Cataloging the Visible Universe through Bayesian Inference at Petascale.* International Parallel and Distributed Processing Symposium (IPDPS), 2018. https://arxiv.org/abs/1801.10277

Dong Yin, Ashwin Pananjady, **Max Lam**, Dimitris Papailiopoulos, Kannan Ramchandran, Peter Bartlett. *Gradient Diversity Empowers Distributed Learning: Convergence and Stability of Mini-batch SGD.* Artificial Intelligence and Statistics (AISTATS), 2018. https://arxiv.org/abs/1706.05699

Kangwook Lee, **Maximilian Lam**, Ramtin Pedarsani, Dimitris Papailiopoulos, and Kannan Ramchandran. *Speeding up Distributed Machine Learning Using Codes.* IEEE Transactions on Information Theory (IEEE Trans. Inf. Theory), 2017. https://arxiv.org/abs/1512.02673

Xinghao Pan, **Maximilian Lam**, Stephen Tu, Dimitris Papailiopoulos, Ce Zhang, Michael I. Jordan, Kannan Ramchandran, Chris Re, and Benjamin Recht. *Cyclades: Conflict-free Asynchronous Machine Learning.* Advances in Neural Information Processing Systems (NeurIPS), 2016. https://arxiv.org/abs/1605.09721

EXPERIENCE

Meta - Research Scientist Intern, FAIR Jun 2022 - Sep 2022

- Designed high-performance GPU algorithms for accelerating private information retrieval for on-device recommendation models.
- Implemented a GPU kernel for distributed point functions, delivering > 1,500× speed-up over an optimized CPU baseline.
- Published results in *ASPLOS 2024*.

Harvard - Graduate Student Researcher, Edge Computing Lab Aug 2019 - May 2024

- Designed *PrecisionBatching*, a GPU-accelerated pipeline for low-batch neural network inference using weight and activation quantization; achieved high efficiency on NVIDIA T4 hardware.
- Co-developed *QuaRL*, applying quantization techniques to reduce communication and energy overhead in distributed multi-node reinforcement learning training.
- Collaborated with MLCommons on *People's Speech*, contributing to data collection, quality control, and evaluation for a large-scale open-source English speech recognition dataset.

- Investigated deployment strategies for on-device machine learning on resource-constrained microcontrollers using quantization, sparsification, and model compression.
- Conducted research on privacy-preserving ML systems, including federated learning, secure multiparty computation, homomorphic encryption, and GPU-accelerated private information retrieval.

Stanford - Graduate Student Researcher, Hazy Research Aug 2018 - May 2018

- Contributed to DARPA D3M's Model Search program, developing efficient neural architecture search and AutoML pipelines for generalized ML problem-solving.
- Leveraged developer log data as a meta-dataset to model and predict deep learning training metrics, such as convergence time and compute cost.
- Investigated quantization techniques for producing compact, efficient word embeddings optimized for memory and compute constrained environments.

Google – Software Engineering Intern, Platforms Team May 2017 – Aug 2017

- Contributed to TensorFlow's internal performance modeling tools (Grappler/TFSim), simulating execution time of computation graphs on hypothetical hardware, including TPU.
- Modeled performance of neural machine translation (NMT) workloads on proprietary topologies and HPC clusters (Intel Xeon + NVIDIA Tesla Volta/Pascal/Kepler GPUs, TPUs).
- Extended early-stage estimator into a production-ready tool supporting generalized HPC configurations and TPU deployment modeling.
- Designed a queuing-theoretic algorithm to forecast distributed training latency for TensorFlow-based NMT.
- Benchmarked inference and training performance across heterogeneous hardware setups.
- Delivered a Google Tech Talk to 100+ engineers, presenting results and tool integration strategy.

Google – Software Engineering Intern, Search Team May 2016 – Aug 2016

- Integrated music genres from Google's Metajam database into the Music Knowledge Graph and Search Knowledge Panels.
- Computed and analyzed genre coverage metrics surfaced by Google Search.
- Optimized load balancing in the Knowledge Graph pipeline, doubling triple generation throughput.
- Deployed the full Search stack on a private cluster to evaluate result quality for music genre queries.

LinkedIn – Software Engineering Intern, Tools Team May 2015 – Aug 2015

- Contributed to the development of LinkedIn's internal code search platform used company-wide.
- Designed and integrated a regular expression search feature, enabling advanced pattern queries.
- Boosted regex search performance by 2–5× using trigram indexing techniques.
- Deployed the enhanced code search engine and regex capability across the engineering organization.

TEACHING EXPERIENCE

Teaching Assistant, Harvard CS242 (Computing at Scale) Spring 2021
Teaching Assistant, Stanford CS107 (Computer Organization & Systems) May 2017 - Aug 2019

AWARDS, HONORS

Winner of The Joint Communications Society & Information Theory Society Paper Award 2020

- Paper: Speeding up distributed machine learning using codes, Kangwook Lee, Maximilian Lam, Ramtin Pedarsani, Dimitris Papailiopoulos, Kannan Ramchandran (IEEE Transactions on Information Theory 64 (3), 1514-1529, 2017)

Theodore H. Ashford Graduate School of Arts and Sciences (GSAS) Fellowship 2019

- Ashford Family support for a cohort of up to six exceptional incoming graduate students at the Harvard Graduate School of Arts and Sciences (GSAS).

SKILLS

python, c, c++, PyTorch, Tensorflow, CUDA, cryptography, parallel/distributed computation, machine learning, machine learning optimization, multiparty computation, differential privacy, quantization, performance optimization, distributed machine learning, private machine learning