Maximilian Lam

OBJECTIVE

Scaling machine learning and artificial intelligence systems towards mass adoption through a combination of systems, infrastructure, hardware, co-design and privacy-preserving/cryptographic 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, Profesor 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

- Researched high performance GPU algorithms for speeding up private information retrieval techniques for on-device machine learning for recommendation
- Developed GPU kernel for a distributed point functions, a critical cryptographic primitive in private table lookups attaining >1,500× speedup over an optimized CPU baseline
- Research paper published in ASPLOS 2024

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

- PrecisionBatching - GPU acceleration for low-batch neural network inference with weight and activation quantization; accelerated on a NVIDIA T4 GPU
- QuaRL - Applying quantization to make distributed multi-node reinforcement learning training more efficient and sustainable
- Peoples Speech - Collaboration with MLCommons on large scale speech dataset construction, including collecting and evaluating audio samples
- Research towards deploying on-device machine learning on tiny powered devices like microcontrollers using techniques like quantization, sparsification

- Research on accelerating privacy preserving machine learning techniques including federated learning, multiparty computation, secure neural network inference and private information retrieval

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

- DARPA D3M Model Search - Research on efficient neural architecture search and AutoML for general problem solving
- Research using developer log data as datasets for model search towards evaluating and predicting training metrics (i.e: training cost) for deep learning
- Research on quantization for efficient and compact word representations

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

- Contributed towards TensorFlow. Specifically, considerably contributions towards an internal TensorFlow performance "estimator" / performance-modeler (Grappler / TFSim) that models and predicts the time taken to run a TensorFlow computation graph on hypothetical hardware devices and hardware device configurations; i.e: TPUv4
- Modeled TensorFlow performance of neural machine translation training on: proprietary (Google) topology, a HPC multi-node cluster of Intel-Xeon NVIDIA NVLINKed Tesla-class Volta-100 / Pascal-100 / Kepler-80 / Kepler-40 GPU / TPU Nodes
- Enhanced and productized early stage implementation of TensorFlow's performance estimator into a working application by enabling performance modeling of running TensorFlow graphs on generalized HPC hardware models, including TPUs
- Developed queuing theory algorithm to estimate performance of distributed machine learning running on TensorFlow infrastructure for neural machine translation (NMT) training
- Models predicted performance of executing neural machine translation and convolutional neural network inference on disparate hardware devices and hardware configurations
- Presented productized TensorFlow performance estimator in Google tech talk to over 100 people

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

- Enhanced Music Knowledge Graph / Knowledge Panels in Google Search application with music genres from Google's Metajam database
- Calculated music genre coverage statistics displayed by the Google Search engine
- Tuned knowledge graph pipeline load balancer and improved performance of knowledge graph triples generation by 100
- Launched Google search stack on private cluster to evaluate quality of new music genres query results

Linkedin - Software Engineering Intern, Tools Team May 2015 - Aug 2015

- Various contributions to Linkedin's internal codesearch tool
- Designed and integrated regular expression search feature to company-wide codesearch tool
- Improved performance of regex searches by 2x-5x by using trigram index algorithms
- Deployed internal codesearch engine and regex feature company wide

TEACHING EXPERIENCE

Teaching Assistant, Harvard CS242 (Computing at Scale)	Spring 2021
Teaching Assistanct, 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, quantization, distributed machine learning, private machine learning