Maximilian Lam

Education

Harvard University *2019 - Now*
PhD in Computer Science, Expected Graduation: June 2024
Committee: Michael Mitzenmacher, Edward Suh, Vijay Janapa Reddi, Gu-Yeon Wei, David Brooks
Advisors: Vijay Janapa Reddi, Gu-Yeon Wei, David Brooks

Stanford University *2017 - 2019*
M.S. in Computer Science

University of California, Berkeley *2013 - 2017*
B.A. in Computer Science
Graduated with High Distinction

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.* 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.* Arxiv Preprint. `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 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 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`

Maximilian Lam*, Sharad Chitlangia*, Srivatsan Krishnan*, Zishen Wan, Aleksandra Faust, Vijay Janapa Reddi. *Quantized Reinforcement Learning (QuaRL)*. Arxiv Preprint, 2020. `https://arxiv.org/abs/1910.01055`

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`

Work Experience

Meta / Facebook – FAIR, Research Scientist Intern *June 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 >1,500x speedup over an optimized CPU baseline
- Research paper published in ASPLOS 2024

Harvard Research – Computer Architecture Group *Aug 2019 - Current*

Stanford Research – Hazy Research *June - Sep 2018*

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

- Worked on Tensorflow internals, specifically on internal performance "estimator" that predicts the time it takes to run a Tensorflow computation graph on particular hardware devices
- Modeled TensorFlow performance of neural machine translation training on: proprietary (Google) topology, a HPC multi-node cluster of /Intel-Xeon NVIDIA NVLINK'ed Tesla-class Volta-100/Pascal-100/Kepler-80/Kepler-40 GPU/TPU Nodes
- Enhanced 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
- Models predicted performance of executing neural machine translation and convolutional neural network inference on disparate hardware devices
- Presented TensorFlow performance estimator in Google tech talk to over 100 people

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

- Worked on incorporating and improving Music Knowledge Graph / Knowledge Panels into Google search
- Enhanced music knowledge panels with music genres from the Google 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 entire Google search stack on private cluster to evaluate quality of new music genres query results

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

- Worked on Linkedin's internal codesearch tool
- Designed and integrated new 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

Teaching Assistant for Harvard CS242 (Computing at Scale) *Spring 2021*

Course on parallel and distributed computing and other systems optimizations applied to modern machine learning techniques. Held office hours and developed course materials.

Teaching Assistant for Stanford CS107 (Computer Organization & Systems) *2017-2019*
The third course in computer science curriculum for undergraduates. Explores introductory topics in computer systems and low level programming with C programming, x86 assembly and gdb debugging. Held office hours and led sections.

Awards & Honors

- The Joint Communications Society/Information Theory Society Paper Award, 2020.
- Harvard University, Ashford Fellowship, 2019.

Github

`https://github.com/agnusmaximus`

Google Scholar

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

Skills

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

Email

agnusmaximus@gmail.com