Samuel J. Kaufman

kaufmans@cs.washington.edu • http://samk.name

EDUCATION

Ph.D., Computer Science & Engineering *in progress*
Expected Graduation: Summer 2026
University of Washington, Seattle
Advised by Rastislav Bodik & René Just

Dissertation contributes a dynamic-programming-based DNN compiler which
- jointly and completely decides tiling, fusion, bufferization, data layout, and re-computation;
- reduces time and space requirements by 10x and 10,000x respectively with a novel memoization table; and
- synthesizes x86 matrix-multiplication and softmax kernels which match or beat the performance of the best available library implementations (OpenBLAS, Intel MKL, PyTorch, etc.).
  - Merged one such kernel into gemma.cpp.

M.S., Computer Science & Engineering
University of Washington, Seattle, March 2020

B.S. in Informatics *emphasis in Human-Computer Interaction*
University of California, Irvine, 2006—2010
Honors Thesis: "Automatic programming with reuse-informed search"
Awarded Chancellor's Award for Excellence in Undergraduate Research

EXPERIENCE

Research Intern, Google Brain — 2019–2020
Developed an ML-based performance model for XLA tensor programs targeting Google's TPU. This work was deployed to production, yielding 10–20% execution speedups and saving ~2% of the total TPU compute time in Google data centers overall.

Co-Founder, Gradient — 2013–2017
Worked with clients to design and build digital products, especially for mobile devices and data-heavy domains. Hired engineers, designers, and sales staff. Developed sales, product, and content strategies.

SELECTED PUBLICATIONS

S. J. Kaufman, R. Just and R. Bodik. "Morello: compiling fast neural networks with dynamic programming and spatial compression." arXiv Preprint. May 2025.

S. J. Kaufman, R. Featherman, J. Alvin, B. Kurtz, P. Ammann and R. Just. "Prioritizing mutants to guide mutation testing." ICSE 2022.

S. J. Kaufman, P. M. Phothilimthana, Y. Zhou, C. Mendis, S. Roy, A. Sabne and M. Burrows. "A learned performance model for Tensor Processing Units." MLSys 2021. March 2021.

G. Fedyukovich, S. J. Kaufman and R. Bodik. "Learning inductive invariants by sampling from frequency distributions." *Formal Methods in System Design*, vol. 56, issue 1–3, pp. 154–177. November 2020.

S. J. Kaufman, P. M. Phothilimthana and M. Burrows. "Learned TPU cost model for XLA tensor programs." ML for Systems workshop, NeurIPS 2019.

P. M. Phothilimthana, A. S. Elliott, A. Wang, A. Jangda, B. Hagedorn, H. Barthels, S. J. Kaufman, V. Grover, E. Torlak and R. Bodik. "Swizzle inventor: data movement synthesis for GPU kernels." ASPLOS 2019.