Lucas Liebenwein

linkedin.com/in/lucas-liebenwein/
lucasliebenwein.com

LLM Inference @ Nvidia | CS PhD @ MIT

Education

PhD & SM, Massachusetts Institute of Technology, Cambridge, MA — 09/16 – 08/21

Computer Science and Artificial Intelligence, GPA: 5.0/5.0

PhD Thesis: Efficient deep learning: from theory to practice.

SM Thesis: Contract-based safety verification for autonomous driving.

B.Sc., ETH Zurich, Switzerland — 09/12 – 08/15

Mechanical Engineering, Major: Robotics and Control

GPA 5.86/6.00 (with highest distinction), Valedictorian

Thesis: Autonomous pairing of distributed flight array modules.

Relevant Experience

Nvidia — New York, NY

Tech Lead — 02/23 - Now

- Engineering lead for TensorRT LLM AutoDeploy (docs, blog), a compiler-driven workflow for converting off-the-shelf PyTorch models into inference-optimized graphs. — 05/25 - Now
- Responsible for driving the roadmap, interacting with the product and engineering leadership for TensorRT, and driving new initiatives.
- Current scope includes kv cache management, speculative decoding, sharding, cuda graph, overlap scheduler, hybrid models, VLMs, and overall compiler architecture.

Engineering Manager

- Lead architect for Model Optimizer, Nvidia's state-of-the-art model optimization library, including NAS, pruning, distillation, sparsity, and quantization. — 02/23 - 05/25
- My team was responsible for researching and publishing recipes for quantization (fp8, nvfp4), sparsity (2:4), and distillation.

OmniML (acquired by Nvidia) — San Jose, CA

Chief Architect & Founding Engineer — 10/21 – 02/23

- Lead the design of the neural network model optimization framework and python package, which is the company's core product used by most customers.
- Supervised a research and engineer for delivering custom edge inference solutions.

Neural Magic (now RedHat) — Somerville, MA

Machine Learning Consultant — 07/21 – 10/21

- Consulted the team on state-of-the-art neural network model optimization techniques.
- Recommended specific actions to improve accuracy & latency (model pruning, NAS, …).

MIT CSAIL — Cambridge, MA

PhD Researcher, Advisor: Prof. Daniela Rus — 09/16 – 08/21

- Research in provable compression methods and neural network model optimization.
- Developed novel verification algorithms for safe motion planning in autonomous driving.

Tesla — Palo Alto, CA

Researcher & Software Engineering Intern, Autopilot — 06/19 – 09/19

- Led the design and development of a lateral motion planner (now part of Tesla FSD).
- Standardized planner testing infrastructure leading to accelerated prototyping.