KILIAN HAEFELI

kilianhaefeli@gmail.com ⋄ +41 78 886 52 11
LinkedIn ⋄ Github ⋄ Google Scholar ⋄ Personal Website

RESEARCH

Neural Representation Learning for Temporal Free Floating Boundary Problems Feb. - July 2023
ETH Zurich and Swiss Data Science Center

- Developed an efficient neural representation model to predict the temporal phase boundaries in 3D printing. Used local convexity and using **Graph Neural Networks** to reduce the model space and induce a smoothing prior.
- **Stack**: **Python**, **PyTorch Lightning**, Poetry, Weights & Biases
- **Supervisor**: PhD Nathanaël Perraudin, PhD Karolis Martinkus & Prof. Roger Wattenhofer

Diffusion Models for Graphs Benefit From Discrete State Spaces Feb. - Sept. 2022

- Developed a **Diffusion Model (DDPM)** for Graphs using Bernoulli perturbations as my Bachelors Thesis.
- Published at the NeurIPS GLFrontiers workshop (**oral**) and the Learning on Graphs Conference Arxiv.
- **Stack**: **Python**, **PyTorch**, **C++**, Networkx, Weights & Biases
- **Supervisor**: PhD Nathanaël Perraudin, PhD Karolis Martinkus & Prof. Roger Wattenhofer

EXPERIENCE

LLM Engineering (Working Student) Oct. 2023 - Jan. 2024
AlephAlpha *Heidelberg, DE*

- Built a RAG application for a customer on top of pre-trained large language models implementing and developing new methods in decoding, fine-tuning and retrieval-augmented generation.
- **Stack**: **Python**, FastAPI, Hugging Face, Docker, Vector Databases (Qdrant)

Junior Data Scientist May - Sep. 2022
Logitech *Zurich, CH*

- Developing, implementing **convolutional deep neural networks** and **LSTM**'s predicting occupancy using stereo-vision cameras and a CO2 time-series.
- **Stack**: Python, NumPy, PyTorch, Lightning.

Junior Developer and Co-Founder May 2020 - May 2022
Airica (formerly eden-senses) *Zurich, ZH*

- Co-founded Airica as part of the technical team, which was successfully sold to Logitech.
- Everything from prototyping, data collection and analysis, and embedded programming to sales and product innovation.
- **Stack**: Python, Bash, C++, Numpy, Scipy, Tableau, PostgreSQL, Node-RED

PROJECTS

Single GPU Attention in CUDA

- From the ground implemented and profiled all kernels for a transformer forward pass in CUDA C. Profiled and analyzed each kernel using Nsight Compute and Systems in order to find bottlenecks in memory throughput and latency hiding: repo.
- **Stack**: **CUDA C**, Pytorch, Nsight Compute, Nsight Systems

EDUCATION

Graduate Student (Exchange), University of Toronto Jan. 2024 - April 2024

- **Coursework:** Statistical Learning Theory, Information Theory, and Parallel Systems

Master Machine Learning and Signal Processing, ETH Zurich 2022 - 2025

- **Coursework:** Advanced Machine Learning, Machine Perception, Optimization (Linear, Combinatorial, Convex), Distributed Algorithms, LLMs, Reliable and Trustworthy AI, Probabilistic AI, Mathematics for Data Science
- **Average Grade:** 5.9/6 (top 2%)

Bachelor of Electrical Engineering and Information Technology, ETH Zurich 2019 - 2022

- **Coursework:** Notably courses on Estimation Theory, Signal Theory, Automata Theory,
- **Average Grade**: 5.62/6