Filip Skogh

Zürich, Switzerland
+41 76 269 4039
`filipskogh99@hotmail.com`
`github.com/fiskrt`

Profile

3+ years of experience in Python and C++. Hackathon enjoyer with over CHF 5000 in prize money. Enthusiastic about writing efficient code related to LLMs and ML. Research experience with Transformers (both encoder and decoder) at IBM Research and video segmentation at ETH Zürich. Knowledge in local inference using `llama.cpp`.

Education

ETH Zürich Exchange student, M.Sc. Computer Science	Sep 2022 - Sep 2023
Chalmers University of Technology M.Sc. Data Science and AI. **GPA: 4.8/5**	Aug 2021 - Sep 2023
Nanyang Technological University Exchange Student, B.Sc. Computer Science	Jan-Jun 2020
Luleå University of Technology B.Sc. Computer Science and Engineering. **GPA: 5.0/5**	Aug 2018 - Jun 2021

Experience

Research Intern IBM Research — Zürich, Switzerland — May 2024 - May 2025

- Meta-learning with hyper-networks, researching encoder-decoder networks that predict weights for other networks.
- ICCV paper in review proposing a new optimal transport normalization layer for Vision Transformers.
- Full research cycle with ideation, literature review, experimentation, and peer-review process.

Junior ML Engineer Logiblox — Zürich, Switzerland — Oct 2023 - Apr 2024

- Increased token generation speed by over 300% for LLM inference with quantization and caching.
- Built compiler that converts visual no-code representation in to executable python code.
- Set up from scratch a dockerized server for inference with CI/CD pipeline.

Master Thesis Computer Vision Lab ETH Zürich — Zürich, Switzerland — Feb-Sep 2023

- Used spatial and temporal features to construct a self-supervised loss function for video segmentation.
- Multi-GPU training on 8xA6000 node with DDP on 100+ GB video datasets.
- Supervised by Prof. Fisher Yu and Dr. Martin Danelljan. `filipskogh.com/thesis.pdf`

Research Intern University of Massachusetts Amherst — Massachusetts, United States — Jun-Sep 2022

- Developed a request scheduler for data-centers that minimize carbon emissions.
- Demonstrated a 70% reduction in carbon emissions without latency degradation.
- Best Paper Runner-up award at the IEEE IGSC 2023 conference in Toronto, Canada.

Software Engineer Intern Orange Cyberdefense — Stockholm, Sweden — Summers 2019 - 2021

- Developed threat response software in Python that block ransomware, C&C servers and take snapshots for forensics. Deployed globally on 50.000+ devices in 70+ countries.

Hackathons and Awards

Modular GPU Kernel Hackathon — AGI House, San Francisco — May 2025

- Built a matrix inversion GPU kernel in Mojo based on the LU-decomposition algorithm.
- Showed a 30x speed-up on MI300X GPUs versus a naive non-parallelized implementation.

Lab 42 Hackathon at the World Economic Forum — Davos, Switzerland — Jan 2025

- Won first prize (CHF 5000) as part of a 3-person team.
- Developed an interactive avalanche segmenter using SAM 2.1, also trained a ResNet model to classify the avalanche type. Predicted the avalanche size using geometric computer vision. All wrapped behind FastAPI with a react front-end.

Projects

Streaming JSON parser: Implemented a C++JSON parser with streaming support at the character level suited for auto-regressive LLMs. Exposed through a clean Python interface using Python bindings.

Packet intercept proxy: C++ project developed continuously for three years by hooking Windows socket API `send` and `recv` using the Detours C++library. Large-scale reverse engineering of project with 100k+ loc. Reverse engineered encryption protocols and ciphers to intercept traffic at packet level.

Neural network certification: Developed custom network layers in PyTorch to propagate intervals through a network allowing us to deterministically prove properties about robustness, certain fairness guarantees and that adversarial attacks are not possible. This project was part of the course Reliable and Trustworthy AI at ETH.

RANSAC: Iterative parameter estimation using RANSAC with optimal hypothesis testing that minimizes the number of tests performed. The project was motivated by the scarcity of available implementation and was based on the original white paper.