Vienna (Austria)

# Elias Frantar

elias.frantar@gmail.com

## Education

**Institute of Science and Technology Austria (ISTA)** **Sep. 2020 - Sep. 2024**

- Ph.D. candidate, Computer Science (Distributed Algorithms and Systems Group), supervised by Dan Alistarh.

**Vienna University of Technology (TU Wien)** **Oct. 2015 - Nov. 2020**

- Dipl. Ing. (equivalent M.Sc.) Computer Science (Logic & Computation), Nov. 2020. GPA: 3.97/4.00
- B.Sc. Computer Science (Software & Information Engineering), Sep. 2018. GPA: 3.98/4.00

*Awards:* Austrian State Prize 2021 awarded to the top 50 out of around 16000 master's graduates each year. Selected for TU Wien's computer science excellence scholarship 4 years in a row, 2016 - 19.

## Experience

HTTPS://GITHUB.COM/EFRANTAR

**ISTA** **Ph.D. Candidate** **Sep. 2020 - Present**

- Developed *GPTQ* and *SparseGPT*, the first quantization and sparsification algorithms fast and accurate enough to successfully compress 100+ billion parameter models to $\leq$ 4-bit precision and $\geq$ 50% sparsity, respectively.
- Implemented the first open-source GPU kernel to accelerate LLM inference via weight-only quantization, as well as *Marlin*, the first int4$\times$fp16 kernel to achieve near-ideal $\approx 4\times$ speedups up to batchsizes of 16 - 32.
- Built *QMoE*, the first framework to compress a trillion-parameter Mixture-of-Experts model to $<$ 1-bit per parameter, in a custom encoding scheme co-designed with an efficient CUDA inference kernel.
- Introduced and refined various second-order-based pruning methods to improve over best-known results across vision and language models, in both cheap post-training and expensive training-based settings.
- Created new techniques for hardware-aware sparsification, with state-of-the-art speed-vs-accuracy trade-offs.

**Google DeepMind** **Student Researcher** **Dec. 2023 - Present**

- Working on new approaches for optimizing the shape and structure of large generative models to maximize inference hardware utilization, especially in low-latency settings with compression and sharding.

**Google Brain** **Student Researcher** **Dec. 2022 - Jul. 2023**

- Performed a scaling study, training a large number of Transformer models on massive vision and text datasets, and identified the first scaling laws for parameter sparsity in the context of foundation models.
- Redesigned state-of-the-art post-training quantization and pruning methods for effective scaling to 100+ hardware accelerators (TPUs) and applied them to compress extremely large neural networks like PaLM-540B.

**TU Wien / ISTA / BASF** **Internships** **Mar-Aug 20 / Oct-Dec 19 / Mar-Aug 18**

- Worked on multiple object tracking, adversarial machine learning, and anomaly detection in time-series.

## Research Impact

HTTPS://SCHOLAR.GOOGLE.COM/CITATIONS?USER=HJDLWZ8AAAAJ

- 6 first-author and 5 co-author major conference papers; overall 500+ citations and 2.5k+ GitHub stars.
- Oral at ICML23 and Spotlight at ICLR24; as well as invited talks at Apple, Google and Amazon.
- SparseGPT covered by Communications of the ACM and national television.
- GPTQ integrated into libraries like HuggingFace, TensorRT-LLM (NVIDIA) and neural-compressor (Intel).

## Projects

HTTPS://YOUTUBE.COM/ELIASFRANTAR

10+ million views on YouTube; presented live at the Royal Institution Christmas Lectures 2019, shown on BBC.

**Cuboth & SquidCuber:** *The world's fastest (unmodified 3x3 / Lego) Rubik's Cube solving robots.*

- Introduced various new techniques, tricks and optimizations to beat 5-year-long standing records being 2$\times$ faster with similar hardware, and achieve major milestones (e.g., first ever $<$ 1s solve with a Lego machine).

**rob-twophase & qphase:** *The current state-of-the-art Rubik's Cube solving algorithms.*

- Established several novel algorithmic ideas to efficiently integrate complex robot mechanics directly into the search process and thus find solutions that are $>$ 20% quicker to execute physically.

**Languages / Technologies:** Python (torch, JAX); CUDA; C++; git; Linux; LaTex

**Publications**

**QMoE: Practical Sub-1-Bit Compression of Trillion-Parameter Models** **Preprint 2023**
*Elias Frantar*, Dan Alistarh

**Scaling Laws for Sparsely-Connected Foundation Models** **Spotlight – ICLR 2024**
*Elias Frantar*, Carlos Riquelme, Neil Houlsby, Dan Alistarh, Utku Evci

**SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot** **Oral – ICML 2023**
*Elias Frantar*, Dan Alistarh

**GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers** **ICLR 2023**
*Elias Frantar*, Saleh Ashkboos, Torsten Hoefler, Dan Alistarh

**Optimal Brain Compression: […] Accurate Post-Training Quantization and Pruning** **NeurIPS 2022**
*Elias Frantar*, Sidak Pal Singh, Dan Alistarh

**SPDY: Accurate Pruning with Speedup Guarantees** **ICML 2022**
*Elias Frantar*, Dan Alistarh

**M-FAC: Efficient Matrix-Free Approximations of Second-Order Information** **NeurIPS 2021**
*Elias Frantar*, Eldar Kurtic, Dan Alistarh

**SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression** **ICLR 2024**
T. Dettmers, R. Svirschevski, V. Egiazarian, D. Kuznedelev, *E. Frantar*, S. Ashkboos, et al.

**ZipLM: Hardware-Aware Structured Pruning of Language Models** **NeurIPS 2023**
Eldar Kurtic, *Elias Frantar*, Dan Alistarh

**CAP: Correlation-Aware Pruning for Highly-Accurate Sparse Vision Models** **NeurIPS 2023**
Denis Kuznedelev, Eldar Kurtic, *Elias Frantar*, Dan Alistarh

**The Optimal BERT Surgeon: […] Second-Order Pruning for Large Language Models** **EMNLP 2022**
E. Kurtic, D. Campos, T. Nguyen, *E. Frantar*, M. Kurtz, B. Fineran, M. Goin, D. Alistarh

**On the Sample Complexity of Adversarial Multi-Source PAC Learning** **ICML 2020**
Nikola Konstantinov, *Elias Frantar*, Dan Alistarh, Christoph Lampert