LLM From Scratch
At the Frontier of Parameter-Efficient Pre-Training

Project supervision: Kevin Glocker

Large Language Models (LLMs) have rapidly gained popularity in industry and academia through their fluent language generation and general-purpose task solving and reasoning capabilities. However, initially only expensive API access to properietary LLMs hosted by US companies such as OpenAI were available. This restricted companies and governments working with sensitive data or privacy conscious individuals from using LLMs in practice. Furthermore, these models could not be efficiently adapted by companies for their own, specialized use cases, limiting their utility further.

Recent efforts from Meta, Google, Apple and Mistral AI have resulted in orders of magnitude smaller yet similarly or more powerful models that can be deployed locally and fine-tuned for specific needs. However, these "small" models still require considerable resources, with billions of parameters making them costly to host at scale. Recent research suggests that even smaller models with as few as 10 million parameters can still produce coherent English. The feasibility for other languages, such as Swedish, remains unknown. Moreover, scaling this methodology for solving more complex tasks remains as an open challenge.

Improvements in the parameter-efficient LLM space would facilitate model customization and deployment at scale, including directly on edge devices, while reducing the carbon footprint of training and inference substantially.

Your tasks will include:

- Train your own transformer language models using cutting-edge methodology from current research on parameter and data efficient LLMs
- Implementing efficient pre-training code optimized for our high performance compute clusters through benchmarking on key metrics such as Model FLOPs Utilization (MFU). Your work can utilize frameworks and libraries, such as pytorch, deepspeed, and nanotron.
- Integration of MLOps workflows and tools to monitor the training process and track training and evaluation artifacts and results
- Evaluation of your LLM both using intrinsic metrics and benchmarks, including estimating the environmental impact of training and inference by, e.g., computing carbon emissions
- Implement an interactive demo to show off the capabilities of your language model, e.g., by integrating it into a Web UI

Through pre-training models you will then answer a research question about highly parameter-efficient LLMs. Possible options include:

- Investigate whether recent successes of highly parameter-efficient language models can be achieved in Swedish. You will apply state-of-the-art data filtering, processing and synthesis techniques and will train and evaluate models at different scales.
  - Follow up research could include, e.g., instruction tuning or retrieval-augmented generation
- Apply a modular LLM architecture to efficiently learn from a diverse, multi-domain dataset.