MASTER THESIS

“Training costs of LLM”

Niko Gritsch

Elite Graduate Program “Data Science”

LMU Munich, September 2022

Large Language Models and training costs

Niko Gritsch has studied in the Elite Graduate Program “Data Science” at the LMU Munich. In his Master thesis he worked with Large Language Models and their training costs.

Quantization in Large Language Models

Large language models have achieved groundbreaking performance in the field ofnatural language processing, but their huge scale creates high costs for training and inference. Quantization compresses the model or its activations, saving memory, time, and energy at the cost of reduced precision. This thesis presents an overview of quantization approaches in LLMs, and investigates how the quality of a quantized model can be measured. It presents the new first-divergent-token (FDT) metric, which directly compares quantized models with the unquantized baseline, offering advantages over existing measures such as perplexity and accuracy. The thesis further investigates when and how errors in quantized models emerge, and presents the new top-2 uncertainty metric, which can be used to predict errors in quantized models with high recall and good accuracy. Finally, it presents a strategy combining quantized and unquantized models, where the top-2 uncertainty helps to reap the benefits of quantized models while avoiding performance degradation.

More on the Elite Graduate Program:

https://www.elitenetzwerk.bayern.de/start/foerderangebote/elitestudiengaenge/uebersicht-elitestudiengaenge/data-science