Contribution ID: **11** Type: **Academic talk (25min)**

Ä Tännschen, please - Matrix operations at the core of modern Natural Language Processing in OpenGPT-X

Friday, 1 July 2022 17:55 (20 minutes)

The field of Natural Language Processing (NLP) has been undergoing a revolution in recent years. Large-scale language models (LLMs), most notably a series of Generative Pre-trained Transformers (GPTs), exceeded all expectations in benchmark scenarios and real life applications such as text generation, translation, question-answering and summarization.
The engine of the NLP revolution is the so-called attention mechanism, which now allows to process longer sentences without "forgetting"important words. This mechanism is implemented in form of a series of matrix products and lends itself to intense parallelization. The pre-training of transformers requires great computational resources and is one example of the increasing AI workload of large High Performance Computing (HPC) facilities.
OpenGPT-X is a joint effort of 10 partners from science and industry to train and provide access to an open LLM based in Europe in order to guarantee its digital and economic sovereignty. Within this project, the pre-training of the LLM is performed at the Jülich Supercomputing Centre.

Primary author: PENKE, Carolin

Presenter: PENKE, Carolin

Session Classification: Scientific Program Friday