FG Data Engineering Systems - Prof. Dr. Tilmann Rabl

Model Lifecycle Analysis Platform II

(Master Project, Winter Semester 2022)

Deep learning (DL) methods have revolutionized many fields by significantly outperforming previous state-of-the-art approaches. Research in this field is still evolving rapidly, leading to a wide range of DL model architectures for various tasks. In corresponding research papers, models are usually trained on a given platform and are evaluated on accuracy measures. Numbers that are relevant for the full model lifecycle like data loading time, training time, inference latency, or storage consumption are not reported. For a given setting, this makes it hard to decide on the optimal model architecture, framework, and hardware stack.

In the previous master project, we developed a model analysis platform for the training stage of deep learning pipelines. For a set of image recognition model architectures, we looked into their training behavior based on the architecture and development framework. One of the insights of the project was the model's footprint and performance dependence on the development framework and execution mode.

The choice of framework not only affects the machine learning pipeline during the training stage, but throughout the whole model lifecycle. The goal of this project is to extend our analysis on the stages of online preprocessing and model deployment in order to make informed decisions on the preprocessing pipeline, serving framework, and used hardware.

Before data is fed into a deep learning model for training or inference it is cached and runs through multiple transformations. The use of transformations and caching differs across pipelines the implementation is highly dependent on the used framework. The impact of these differences is reflected in the overall resource consumption, convergence, and the performance of the model. For this aspect, subjects of interest are the data loading latency, resource utilization, pipeline staleness, and the overall size of cached and intermediate data.

After training and testing a model it is deployed to one or multiple inference servers to make predictions on real data. For this stage of the model lifecycle, important aspects are the initial deployment of the model to the servers, handling of inference requests, and updating deployed models. Based on the model architecture and serving framework, the metrics of interest are the deployment cost, inference latency, concurrent model access latency, and the cost of any model updates/redeployments.

You will become familiar with different deep learning model architectures and their training and inference behavior. You will have the opportunity to build a model lifecycle analysis platform that will measure and store key model and pipeline characteristics. The project will be implemented in Python.

Contact

Ilin Tolovski, Nils Straßenburg, Ricardo Salazar Diaz

Grading

Courses applicable: ITSE (Masterprojekt), DE (Data Engineering Lab)

Graded activity:

- Implementation / group work
- Final report (8 pages, double-column, ACM-art 9pt conference format)
- Final presentation (20 min)