Model Lifecycle Analysis Platform

(Master Project, Summer 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 larger models with more complex architectures. This results in a wide range of DL model architectures for a variety of tasks. The evaluation of the newly developed models focuses on the model prediction performance, like the top-1 and top-5 error for the ImageNet challenge [1]. Models are usually developed and evaluated on one platform only, and metrics like inference/training time, model storage consumption, and deployment costs are not taken into account. Thus, the evaluation of the models does not consider the entire model lifecycle but only the initial training/model development. This makes it hard to decide for a model architecture and a training/deployment system when there are certain constraints on, for example, training time, inference latency, or storage consumption.

The goal of this project is to develop a platform that provides measurement information about end-to-end model lifecycles on a set of representative deep learning tasks across different domains and platforms. Each model's lifecycle is represented by a model pipeline that consists of two aspects: (1) The model development, which includes the definition of the model architecture, and the (re-)training of a model. The metrics of interest for this aspect are the model training time, the model performance, the model storage consumption, and the save/recovery time. (2) The model deployment, which includes the initial deployment of a model, serving inference requests, and any subsequent redeployments/updates. For this aspect, relevant measures are the model deployment cost, inference time/latency for given batch sizes, and the costs of model updates and redeployments.

You will become familiar with different DL 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 characteristics. The project will be implemented in Python.

Contact

Ilin Tolovski, Nils Straßenburg

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)

References

[1] ImageNet Challenge: https://www.image-net.org/challenges/LSVRC/