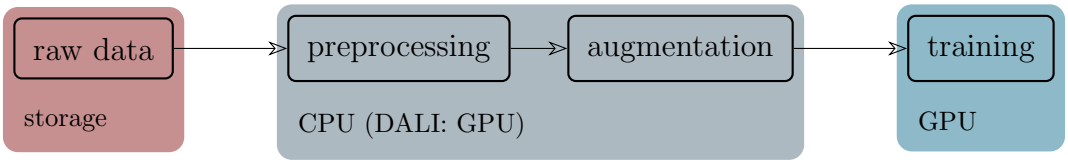
PyTorch NVIDIA DALI DYNAMICS GROUP DYN

Project Thesis

Implementation of a fast and flexible data pipeline for cutting-edge machine learning performance

The training of machine learning (ML) models relies on large amounts of training data to achieve competing performance. Big data applications require data pipelines which establish a constant data flow by loading small subsets directly from the disk, hence bypassing memory limitations. Thereby, the reduction of GPU idle times by means of efficient I/O and preprocessing is a key challenge to accelerate ML training.



This project thesis aims at the implementation of a robust and flexible data pipeline for PyTorch using state of the art approaches like NVIDIA Data Loading Library (DALI) or PyTorch LightningDataModule. The efficient use of available resources implies a fast data format like TFRecord or protobuf to reduce I/O times. Furthermore, online preprocessing promotes flexibility and reduces the storage footprint of the data to a minimum.

The scope of this work covers the following tasks:

- Literature review of the state of the art,
- Implementation of high performance data pipeline and
- Validation against existing approach based on Tensorflows tf.data API.

Prerequisites:

- Demonstrated programming experience in Python,
- General knowledge about machine learning and data handling,
- Curiosity, excellent skills in independent work and communication.

Contact: M.Sc. Mathies Wedler (mathies.wedler@tuhh.de)

Dynamics Group (M-14)
Hamburg University of Technology
www.tuhh.de/dyn