Novelty-based data selection on the edge for efficiently training deep neural networks in the cloud

Type: Bachelor, Master

Contact
IMA: Christian Scheiderer | Tel 0241 8091139 | christian.scheiderer@ima-ifu.rwth-aachen.de
LLT: Moritz Kröger | Tel: 0241 8906670 | moritz.kroeger@llt.rwth-aachen.de

Scientific Question

Deep neural networks are successfully employed in various production processes for finding and learning correlations between process parameters and quality measurements. While most of the time a lack of data is the bottleneck for machine learning projects, other processes produce data in volumes, which are infeasible to transmit from the edge to central storage infrastructure. One such process is UKP (Ultrakurzpuls / ultrashort pulse) lasercutting, where multiple gigabytes of data are produced every minute. As not all data is equally valuable for training machine-learning models, it becomes clear that methods for effectively filtering samples with respect to their expected benefit for training promise the potential in terms of data efficiency. One first approach in this direction is given by novelty-detection algorithms, which are able to identify sparsely occurring data samples. The scientific question to be answered by this thesis is focused on investigating, to what extend novelty-detection algorithms can be employed for improving sample efficiency in the training of deep neural networks.

Scientific methodology

As a first step, a familiarization with the UKP process and the domain of supervised machine learning is conducted. Subsequently, a machine learning pipeline is constructed which allows training models using a dataset from the UKP process. After an investigation of the effect of the sample order during training, methods suitable for identifying novelty of data samples are implemented and thoroughly evaluated.

Aim and expected results

This thesis aims at investigating to which extend novelty-based machine learning methods suitable for robustly predicting the benefit of training samples in order to realize prioritization and filtering of data on the edge.