Contribution ID: **158** Type: **Presentation**

Deep Learning for Predicting the Popularity of Datasets

Wednesday, 10 October 2018 16:45 (15 minutes)

Accessing datasets stored on tape drives is comparatively time-consuming. Therefore, a certain fraction of all datasets is usually provided on a cache storage built of hard disks. Caching algorithms are used to identify popular datasets and to move them in advance from tape drives to the cache storage. In general, there is a considerable gap between the effectiveness of traditional caching algorithms and the optimal (or Belady) caching algorithm. It seems to be unlikely that the gap can be reduced significantly by optimizing traditional caching algorithms. The aim of our project is to explore whether popular datasets can be identified more optimally by applying deep learning methods.

Training a neutral network is time-consuming. This is true, in particular, if the training sets are large. The Atlas experiment at the Large Hadron Collider (LHC) stores every access to datasets in log files (many parameters are saved such as the name of the file, name of the dataset the file belongs to, the tool used for accessing the file, and the access time). In total log data of the order of 0.5 TB are stored per month. Applying deep learning techniques to large datasets needs a scalable infrastructure. To speed up the training of neural networks, several proposals were submitted, for example the use of specialized processors like GPUs or TPUs. We designed a cluster of containers for running neural networks in parallel. The cluster allows to investigate different distributed deep learning strategies, e.g. data parallelism and model parallelism. To distribute files across the nodes of the cluster and to train neural networks in parallel, the big data analytics frameworks Apache Flink and Apache Spark are used.

The talk gives an overview of the current status of our project. The machine learning workflow running on the cluster system is presented. First results obtained by applying a Convolutional Neural Network to a small subset of Atlas log data are shown. The speedup of different parallelization strategies is evaluated. An outlook on ongoing work will be given.

Type of abstract

Presentation

Summary

There are many “traditional” caching algorithms for solving the problem of determining the popularity of datasets. Our project explores whether popular datasets can be identified more optimally by applying deep learning methods. We have developed a cluster of containers for comparing different distributed deep learning strategies. The talk gives an overview of the current status of our project and presents first results.

Primary author: Mrs ZIMMERMANN, Nina (Univ. of Applied Sciences (HTW) Berlin)

Co-authors: Mr NAGEL, Daniel (Univ. of Applied Sciences (HTW) Berlin); Mr THOM, Florian (Univ. of Applied Sciences (HTW) Berlin); Mr FUCHS, Hannes (Univ. of Applied Sciences (HTW) Berlin); Prof. HESSLING, Hermann (Univ. of Applied Sciences (HTW) Berlin); Mr STRUTZ, Marco (Univ. of Applied Sciences (HTW) Berlin); Mr MENZEL, Maximilian (Univ. of Applied Sciences (HTW) Berlin); Mr WOCHINGER, Tobias (Univ. of Applied Sciences (HTW) Berlin)

Presenter: Mrs ZIMMERMANN, Nina (Univ. of Applied Sciences (HTW) Berlin)

Session Classification: Computing Services Part II

Track Classification: Area 3. Computing and Virtual Research Environments