DT-DB4MLKD-B: Moderne Datenbanksysteme für maschinelles Lernen und Wissensentdeckung
DT-DB42-M: Database Systems - The Question to or the Better Answer on 42?

Maximilian E. Schüle

19. April 2023

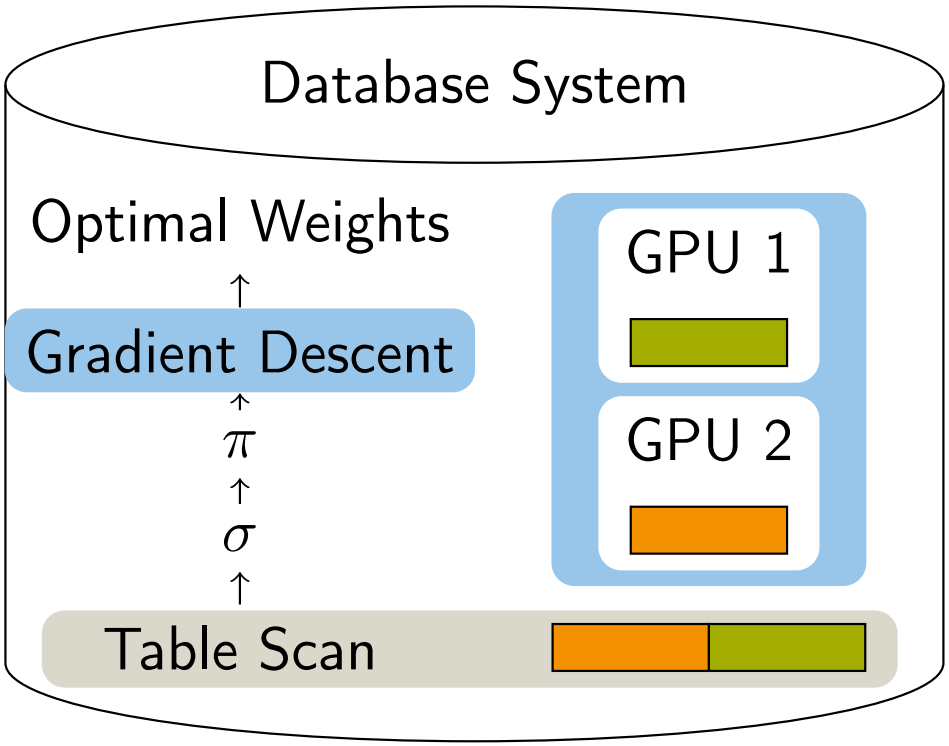
mailto:maximilian.schuele@uni-bamberg.de

Goal

In this seminar, we study the challenges of modern database systems. We discuss the topic along with very recent publications about database systems for machine learning and knowledge discovery.

- paper discussion group
- your task:
  - pick one paper
  - write four pages (double-column, ACM style)
  - presentation (20 minutes + 10 minutes discussion)

In-Database Machine Learning



- SQL sufficient for machine learning (ML)
  - Turing-complete with recursive tables
- Idea
  - Data preprocessing using SQL
  - No need for data extraction out of a database system
  - Continuously train models using operators for gradient descent with GPU support
  - Label data within the database system using SQL

Database Conferences

- VLDB `https://vldb.org/pvldb/volumes/`
- SIGMOD `https://dl.acm.org/sig/sigmod`
- ICDE `https://ieeexplore.ieee.org/xpl/conhome/1000178/all-proceedings`
- EDBT `https://www.openproceedings.org`
- SSDBM `https://dl.acm.org/conference/ssdbm`
- ADBIS `https://adbis2022.polito.it/proceedings/`
- BTW `https://dl.gi.de/handle/20.500.12116/35791`

Timeline

- We 14-16 Uhr WE5/05.005
- 19.04. (today)
- 17./24.5./7./14./28.6. (5./12.7.) presentation slots
- send e-mail with your topic wish until 1.5. to `mailto:maximilian.schuele@uni-bamberg.de`
- 2 weeks prior to your presentation: structure of your draft
- 1 day before your presentation: send me your slides
- 1 week after your presentation: send me your paper

Selected Papers for Discussion

- ML
  - Enabling SQL-based Training Data Debugging for Federated Learning p388-wu.pdf
  - A Scalable AutoML Approach Based on Graph Neural Networks (p2428-helali.pdf)
  - UPLIFT: Parallelization Strategies for Feature Transformations in Machine Learning Workloads (p2929-phani.pdf)
  - Optimizing Machine Learning Inference Queries with Correlative Proxy Models p2032-yang.pdf
  - Exploration of Approaches for In-Database ML `https://openproceedings.org/2023/conf/edbt/paper-7.pdf`
  - Accelerating Python UDFs in Vectorized Query Execution `https://www.cidrdb.org/cidr2022/papers/p33-klaebe.pdf`
  - Learned Selection Strategy for Lightweight Integer Compression Algorithms `https://openproceedings.org/2023/conf/edbt/3-paper-48.pdf`
  - Enhanced Featurization of Queries with Mixed Combinations of Predicates for ML-based Cardinality Estimation. `https://openproceedings.org/2023/conf/edbt/paper-1.pdf`

- DBMSes
  - YeSQL: "You extend SQL" with Rich and Highly Performant User-Defined Functions in Relational Databases p2270-foufoulas.pdf
  - Containerized Execution of UDFs: An Experimental Evaluation p3158-saur.pdf
  - Hardware Acceleration of Compression and Encryption in SAP HANA p3277-chiosa.pdf
  - SQLite: Past, Present, and Future p3535-gaffney.pdf
  - DBOS: A DBMS-oriented Operating System p21-skiadopoulos.pdf
  - Evaluating Query Languages and Systems for High-Energy Physics Data p154-muller.pdf
  - Patched Multi-Key Partitioning for Robust Query Performance `https://openproceedings.org/2023/conf/edbt/paper-13.pdf`
  - A simplified Architecture for Fast, Adaptive Compilation and Execution of SQL Queries https://www.openproceedings.org/2023/conf/edbt/paper-156.pdf

Selected Papers for Discussion (con'd)

- GPU
  - Orchestrating Data Placement and Query Execution in Heterogeneous CPU-GPU DBMS (p2491-yogatama.pdf)
  - Efficient Load-Balanced Butterfly Counting on GPU (p2450-zhang.pdf)
  - Harmony: Overcoming the Hurdles of GPU Memory Capacity to Train Massive DNN Models on Commodity Servers (p2747-li.pdf)

- Matrix algebra
  - Query Processing on Tensor Computation Runtimes (p2811-he.pdf)
  - Share the Tensor Tea: How Databases can Leverage the Machine Learning Ecosystem p3598-interlandi.pdf
  - Federated Matrix Factorization with Privacy Guarantee p900-li.pdf
  - Improving Matrix-vector Multiplication via Lossless Grammar-Compressed Matrices p2175-tosoni.pdf
  - Density-optimized Intersection-free Mapping and Matrix Multiplication for Join-Project Operations p2244-chen.pdf

Writing Papers

In-Database Machine Learning with SQL on GPUs

Maximilian E. Schüle, Harald Lang, Maximilian Springer, Alfons Kemper, Thomas Neumann, and Stephan Günnemann
{m.schuele,harald.lang,max.springer}@tum.de,{kemper,neumann,guennemann}@in.tum.de
Technical University of Munich

ABSTRACT

In machine learning, continuously retraining a model guarantees accurate predictions based on the latest data as training input. But to retrieve the latest data from a database, time-consuming extraction is necessary as database systems have rarely been used for operations such as matrix algebra and gradient descent.

In this work, we demonstrate that SQL with recursive tables makes it possible to express a complete machine learning pipeline out of data preprocessing, model training and its validation. To facilitate the specification of loss functions, we extend the code-generating database system Umbra by an operator for automatic differentiation for use within recursive tables: With the loss function expressed in SQL as a lambda function, Umbra generates machine code for each partial derivative. We further use automatic differentiation for a dedicated gradient descent operator, which generates LLVM code to train a user-specified model on GPUs. We fine-tune GPU kernels at hardware level to allow a higher throughput and propose non-blocking synchronisation of multiple units.

In our evaluation, automatic differentiation accelerated the runtime by the number of cached subexpressions compared to compiling each derivative separately. Our GPU kernels with independent models allowed maximal throughput even for small batch sizes, making machine learning pipelines within SQL more competitive.

KEYWORDS

In-Database Machine Learning, Automatic Differentiation, GPU

ACM Reference Format:
Maximilian E. Schüle, Harald Lang, Maximilian Springer, Alfons Kemper, Thomas Neumann, and Stephan Günnemann. 2021. In-Database Machine Learning with SQL on GPUs. In *33rd International Conference on Scientific and Statistical Database Management (SSDBM 2021), July 6–7, 2021, Tampa, FL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3468791.3468840

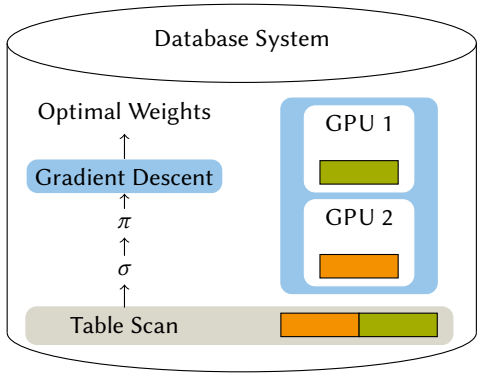


Figure 1: In-database machine learning: gradient descent with GPU support, embedded in a query plan.

training are co-processed on graphical processing units (GPUs) or tensor processing units (TPUs) developed for this purpose.

Integrating machine learning pipelines into database systems is a promising approach for data-driven applications [1, 15, 19, 57, 62]. Even though specialised tools will outperform general-purpose solutions, we argue that an integration in database systems will simplify data provenance and its lineage, and allows complex queries as input. So far, machine learning pipelines inside of database queries are assembled from user-defined functions [13, 30, 41, 54, 65] and operators of an extended relational algebra. This brings the model close to the data source [58] with SQL [2] as the only query language. As modern HTAP main-memory database systems such as SAP HANA [38], HyPer [20, 27, 39, 46] and Umbra [26, 40, 45] are designed for transactional and analytical workload, this allows the latest database state to be queried [24, 43]. But for continuous

Writing Papers: Abstract

- Abstract
  - Rationale
  - Problem
  - Objective
  - Methods / Materials
  - Contributions / Results
  - Conclusion / Implications

Writing Papers: Introduction, Conclusion

- Introduction
  - big picture (Theory, State of the art)
  - Relevance of study field & former studies
  - Research question (Hypothesis)
  - my solution (Approach, Scope, Delimitations, Preview)
  - Scope
  - Delimitations: none?
- Conclusion
  - Review (summary of results)
  - Coupling results to RQ -> validated thesis
  - Conclusions
  - significance of results & conclusions
  - relation to previous research (incl. conflicts)
  - remaining questions
  - recommendations for application & future research -> Outlook

Writing Papers: Related Work, Method, Evaluation

- Related Work
  - officially: proof that you understand the topic
  - inofficially: make reviewers happy
  - sometimes: really interesting study that you used
  - for bibtex entries: https://dblp.uni-trier.de/
- Method
  - nice figures, (pseudo) code, explanation
- Evaluation
  - Comparison to other systems: performance/accuracy/cost/...
- fun: `https://pdos.csail.mit.edu/archive/scigen/`
- ACM template (acmart, sigconf): `https://db.in.tum.de/teaching/ws2122/seminarHauptspeicherdbs/paper-template-suggestion.zip` `https://www.acm.org/publications/proceedings-template`