Samsara: Declarative Machine Learning on Distributed Dataflow Systems

Overview

Motivation:
- apply ML to large datasets stored in distributed filesystems using dataflow engines like Spark

Problem:
- dataflow engines are hard to program:
  - programs consist of sequence of parallelizable second-order functions on partitioned data
  - mismatch for ML applications which mostly operate on tensor type

Samsara
- domain-specific language and execution layer for declarative machine learning
- allows advanced users to rapidly create new algorithms and adapt existing ones
- reduces need for knowledge of programming and execution model of underlying systems

Compilation of Programs to DAGs of Logical Operators

Example: Transpose-Times-Self on Spark

Variant A: Distributed Computation of $X'X$ via Summation of Partial Outer Products

Variant B: Distributed Computation of $X'X$ via Summation of Local Gram Matrices

Choice of Physical Operators

Example: Transpose-Times-Self on Spark

Evaluation, Limitations, Future Work

Benefit of Proposed Optimizations
- compared standard execution mode of Samsara with a variant that has optimizations disabled
- cluster of 24 machines, synthetically generated matrices with 20 columns and a growing number of rows
- 5x performance improvements due to choice of specialized operator for $X'X$ and 3x due to choice of local vs distributed joins in $X'Y$

Future Work & References
- Limitations & Future Work
  - lack of speed for in-core operations due to JVM-based matrix libraries
  - currently exploring integration of ViennaCL, for selected operators to provide a bridge to native performance on many-core architectures
  - high variance in performance between different backends
  - e.g., due to lack of efficient caching of intermediate results in Apache Flink
- Acknowledgements
  - Samsara is the result of a community effort from the Apache Mahout project, with fundamental contributions to design and codebase by Dmitriy Lyubimov
- References
  - “Mahout Scala and Spark Bindings” http://mahout.apache.org/users/sparkbindings
  - “Mahout Scala and Spark Bindings” http://mahout.apache.org/users/sparkbindings