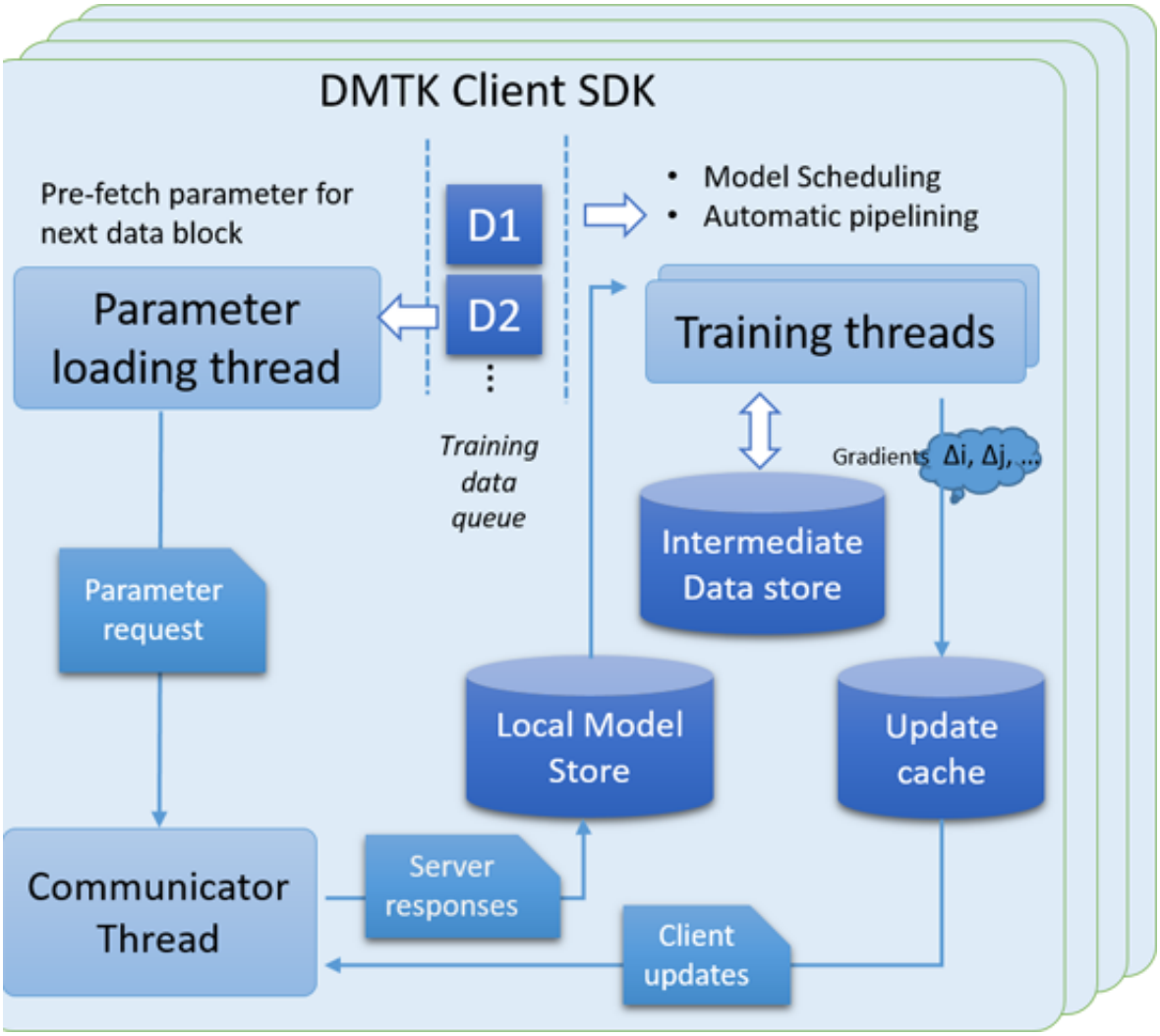
Microsoft open sources Distributed Machine Learning Toolkit for more efficient big data research

November 13 2015



Researchers at the Microsoft Asia research lab this week made the Microsoft Distributed Machine Learning Toolkit openly available to the developer community.

The toolkit, available now on GitHub, is designed for distributed machine learning—using multiple computers in parallel to solve a complex problem. It contains a parameter server-based programing framework, which makes machine learning tasks on big data highly scalable, efficient and flexible. It also contains two distributed machine learning algorithms, which can be used to train the fastest and largest topic model and the largest word-embedding model in the world.

The toolkit offers rich and easy-to-use APIs to reduce the barrier of distributed machine learning, so researchers and developers can focus on core machine learning tasks like data, model and training.

The toolkit is unique because its features transcend system innovations by also offering machine learning advances, the researchers said. With the toolkit, the researchers said developers can tackle big-data, big-model machine learning problems much faster and with smaller clusters of computers than previously required.

For example, using the toolkit one can train a topic model with one million topics and a 20-million word vocabulary, or a word-embedding model with 1000 dimensions and a 20-million word vocabulary, on a web document collection with 200 billion tokens utilizing a cluster of just 24 machines. That workload would previously have required thousands of machines.

In addition to supporting topic model and word embedding, the toolkit also has the potential to more quickly handle other complex tasks involving computer vision, speech recognition and textual understanding.

Specifically, the toolkit includes the following key components:

- DMTK framework: A parameter server, which supports storing a hybrid data-structure model, and a client SDK, which supports scheduling client-side, large-scale model training and maintaining a local model cache syncing with the parameter server side model.
- LightLDA: A new, highly efficient algorithm for topic model training that can process large-scale data and model even on a modest computer cluster.
- Distributed Word Embedding: A popular tool used in natural language processing, the toolkit offers the distributed implementations of two algorithms for word embedding: The standard Word2vec algorithm and a multi-sense algorithm that learns multiple embedding vectors for polysemous words.

In the future, the researchers said, more components will be added to new DMTK versions. Microsoft researchers are hoping that, by open sourcing DMTK, they can work with machine learning researchers and practitioners to enrich the algorithm set and make it applicable to more applications.

More information: More information about the DMTK is available here: www.dmtk.io/index.html

Provided by Microsoft