Combining Big Data Analytics and Deep Learning for the Large Hadron Collider

With a circumference of 27 kilometers and more than 6,000 superconducting magnets, the Large Hadron Collider (LHC) at CERN, the European Organization for Nuclear Research, is the world's largest machine and most sophisticated scientific instrument. The LHC is capable of accelerating protons to 99.9999991% of the speed of light and generating terrifically energetic collisions that are in some cases 100,000x hotter than our sun's core.

These experiments produce stunning amounts of raw data. Up to 1 billion collisions per second take place in the LHC, generating as much as one petabyte (1,024 terabytes or a *million* gigabytes) of data flowing through CERN's systems each second – far more than can be stored by even the world's foremost research institutions. CERN uses proprietary software to filter this data — selecting collision events of interest — before storing it for later analysis.

Seeing an opportunity to improve event filtering using deep neural networks (DNNs) and desiring to shift to open source tools, CERN and a consulting team of Intel engineers used BigDL, Analytics Zoo, and Apache Spark* to develop and test an end-to-end big data analytics/deep learning data pipeline that could potentially address this particularly challenging data problem.

This approach enabled CERN to prototype groundbreaking deep learning algorithms on the same infrastructure it uses for big data analytics. The experiment will improve high-energy physics event filtering techniques, which will improve analysis and save compute and storage resources.

New Approaches to the Biggest Big Data Problems

CERN's current filtering apparatus reduces the data the LHC generates to a more manageable level – one to ten gigabytes per second – a significant savings, but still a lot of data. Improving the filtering system's accuracy for high-energy physics research remains an area worthy of study, as opportunities remain for compute and storage savings.

CERN based its own experimental converged end-to-end data pipeline on the work of T. Nguyen et al., who have proposed the use of DNNs to identify three different event topologies of interest. CERN was particularly interested in this approach's ability to reduce the number of false positives generated by current methods, which often are implemented as complex rule-based systems. Intel-developed open source tools proved essential to CERN's implementation of these algorithms.

Converging Big Data Analytics and Deep Learning

CERN had several requirements for its deep learning infrastructure. Foremost, the solution needed to integrate with CERN's existing Apache Spark big data analytics system, which runs atop a YARN*/Hadoop* cluster and, more recently, Kubernetes*. CERN also desired a system that would scale effectively using CPU compute to increase utilization of existing infrastructure and minimize the cost and complication of additional hardware accelerators and associated programming models. To address these requirements, the CERN openlab selected Intel's open source software (including the BigDL framework and Analytics Zoo toolkit) with support for Python*/PySpark* and standard APIs (specifically Keras* API) for processing the neural network running on their Intel® Xeon® Scalable processor-based systems.

BigDL is a distributed deep learning library for Apache Spark, on which deep learning applications can be written as standard Spark or Python programs, taking advantage of the scalability of an existing Spark cluster. As its name suggests, BigDL makes it easy to integrate deep learning into a big data analytics system. BigDL gave CERN the ability to add deep learning capabilities into their existing infrastructure, creating an end-to-end converged data pipeline.

CERN's adoption of BigDL was facilitated by Analytics Zoo, a unified analytics/AI platform for Spark, TensorFlow*, Keras, and BigDL. Analytics Zoo unites these applications into a single integrated platform that can then be executed atop a Hadoop/Spark cluster for scalable deep learning training or inference. Both BigDL and Analytics Zoo are distributed under the Apache 2.0 license, fulfilling CERN's desire to use open source tools for this project.

Results

The performance of distributed training via BigDL and Analytics Zoo was a key factor in this exercise, as reduced training time means greater productivity for CERN's physicists and data scientists. CERN was pleased with the preliminary scaling performance of the classifier tested. CERN ultimately found that using 20 executors with 6 cores each (for a total of 120 allocated cores) and a batch size of 128 per worker resulted in a training speed of up to 100,000 rows per second, which was sustained throughout the training of the whole dataset.

CERN recognized BigDL's and Analytics Zoo's ability to scale out training using familiar Keras APIs on the Spark cluster already available at CERN. Apache Spark's usage as the main driver of CERN's end-to-end pipeline made the use of BigDL for integrated deep learning all the more suitable. CERN plans future work to further investigate scalability and performance on this infrastructure.

Getting Started with BigDL

We've seen BigDL deliver value in industries like financial services and security, but it's particularly rewarding to see BigDL be used by CERN to investigate some of the ultimate mysteries of science. We look forward to future collaboration with our colleagues at CERN to leverage open source tools and Intel hardware and engineering expertise to solve challenging data problems and enable unprecedented research opportunities.

For more information on CERN's data pipeline for machine learning, refer to their recent blog post. Please follow us on @IntelAI and @IntelAIResearch for more on Intel's efforts to enable the AI and big data communities.

We would like to acknowledge the contributors to this project. The primary contributor is Matteo Migliorini, during his stay at CERN in 2018 under the supervision of Luca Canali. Additional credits go to Viktor Khristenko and Maria Girone of CERN openlab, to Marco Zanetti of Padua University and to the authors of the research paper "Topology classification with deep learning to improve real-time event selection at the LHC", in particular to Maurizio Pierini and Thong Nguyen. Credits to the BigDL and Analytics Zoo team at Intel, in particular, many thanks to Jiao Wang.