REPORT REPRINT

IBM hopes to scale enterprise machine learning with Watson Machine Learning Accelerator

APRIL 8 2019

By Nick Patience, Jeremy Korn

As enterprises move beyond the initial adoption of machine learning and begin to experiment with deep-learning techniques, many find it challenging to deploy and maintain their initiatives. Watson Machine Learning Accelerator is a software package with a variety of open source features that address common barriers to scaling machine-learning workloads.

451 Research®

Introduction

AI and machine learning have made substantial strides in terms of enterprise adoption. In fact, according to data from 451 Research's Voice of the Enterprise: AI & Machine Learning 2H 2018 survey, 20% of respondents have already implemented machine learning in the enterprise, and an additional 20% have a machine-learning initiative in the proof-of-concept stage.

As organizations seek to expand these initiatives, they often run into several barriers, such as the lack of skilled resources, or the complication of deploying and maintaining models. Experimentation with deep learning – a class of machine learning that employs multilayered neural networks to identify abstract structures hidden within large data sets – will only compound these problems. IBM hopes to address some of the most common issues to help usher in the next stage of enterprise adoption.

451 TAKE

IBM brings a deep portfolio of machine-learning tools to the table, from its servers and workbench systems to its Watson Studio development platform. Watson Machine Learning Accelerator, IBM's deep learning-in-a-box system, is a software package containing a variety of features and tools – for example, Large Model Support (LMS), SnapML, Elastic Distributed Training (EDT) and Auto Hyper Parameter Optimization – that seek to make machine learning easier and more scalable for enterprise users. Although Watson Machine Learning Accelerator targets a rarefied group of developers with large workloads and big infrastructure budgets, the product makes strategic sense for IBM, given both the growing importance of deep-learning workloads for AI applications, and the hardware and software expertise IBM can bring to solving the relevant problems.

Context

IBM maintains an extensive machine-learning portfolio, running the gamut from high-throughput hardware systems such as the IBM Power System series, to Watson-enhanced horizontal and vertical software applications, as well as custom AI consulting services. In some sense, IBM Watson Machine Learning Accelerator serves to connect IBM's software and hardware offerings to enable technical users within enterprises to supercharge their machine-learning projects built in Watson Studio, IBM's machine-learning development platform.

Product

Watson Machine Learning Accelerator (previously IBM PowerAI Enterprise) is a combined hardware and software package that aggregates a variety of prominent open source deep-learning frameworks alongside development and management tools, so that enterprise users can more easily build and scale machine-learning pipelines. The software supports the Caffe, TensorFlow, PyTorch and Keras frameworks, and includes a variety of modules such as LMS and SnapML, as well as the EDT and Auto Hyper Parameter Optimization features.

Watson Machine Learning Accelerator integrates open source technologies, and is compatible with other machine-learning tools, so customers can integrate their favorite features from other providers of machine-learning software. IBM attaches Watson Machine Learning Accelerator to sales of its Power System S822LC and AC922 servers, which integrate IBM POWER8 or POWER9 CPUs with NVIDIA Tesla P100 or V100 GPUs to address the HPC needs of enterprises. The software can also run IBM's x86 servers, although not all functionalities are optimized for this type of deployment.

The internal bandwidth of many accelerated servers is a significant hurdle for data scientists looking to train complex deep-learning models using expansive data sets. The LMS functionality of Watson Machine Learning Accelerator leverages Nvidia's NVLink, directly connecting IBM CPU-based hardware with NVIDIA GPUs, providing upward of 5.6x data transfer speeds to system memory. Users can thus tackle projects where model size or data size are significantly larger than the limited memory available on the GPUs, leading to more accurate models and improving model training time. The programming and compute efficiencies realized through LMS are substantial: in one instance, IBM was able to train a model on the Enlarged ImageNet Dataset 3.8x faster than without LMS.

Another significant barrier for deep-learning projects is the substantial time it takes to train models, which can slow development cycles and delay project timelines. To accelerate the training process, Watson Machine Learning Accelerator includes SnapML, a distributed machine-learning library for GPU acceleration supporting logistic regression, linear regression and support vector machine models. These preinstalled models make building and training more efficient. For example, a logistic regression model trained using SnapML and 4 Description POWER9 servers with GPUs was trained 46x faster than when it was trained using 90 x86 servers.

Scaling jobs across compute nodes is another challenge at the leading edge of deep learning. Watson Machine Learning Accelerator seeks to address this obstacle with EDT, a feature that allows users to both distribute jobs across multiple compute nodes and elastically allocate GPU resources. The dynamic scaling enabled by EDT allows researchers to more easily prioritize machine-learning jobs. A second option to scaling is the Distributed Deep Learning (DDL) feature, which allows users to allocate a training job across multiple servers with minimal communication inefficiencies.

Defining and tuning hyper parameters is often a time-consuming and tedious part of the machine-learning process. The hyper parameter optimization feature within Watson Machine Learning Accelerator allows users to automate this process by building and comparing a series of models in parallel.

Finally, Watson Machine Learning Accelerator uses IBM Spectrum Conductor, a Spark-based, machine-learning workload lifecycle manager and scheduler, which helps make sure researchers are utilizing compute resources at their maximum capacity. IBM views Spectrum Conductor as a real differentiator as enterprises manage access to what are still quite rare and expensive compute resources. It has found data scientists overcompensating in terms of blocking CPU or GPU resources, and claims Spectrum can schedule jobs more intelligently based on the nature of the job, helping enterprises scale their machine-learning efforts.

Customers

IBM reports hundreds of paying clients and an extensive pipeline of prospects for Watson Machine Learning Accelerator. Customers span a variety of vertical markets such as financial, healthcare and energy organizations, and range from small AI-focused startups to large multinationals. The common denominator is that users are looking to apply machine learning at scale within their organizations.

BP is one customer using Watson Machine Learning Accelerator to accelerate its deep-learning workloads. The oil and gas multinational's use case centers on 3-D computer vision and seismic processing, which necessitates higher throughput and memory requirements than 2-D applications. By combining DDL and LMS, researchers can scale workloads while maintaining training efficiency. They also reported an overall simplification to process, from job creation to a reduction in code size and easier code maintainability.

Another customer is DeepZen, which uses the software to create more realistic-sounding NLG engine for audiobooks and voiceovers. A third customer is Wells Fargo, which uses the tools within the Watson Machine Learning Accelerator platform to validate hundreds of models a day.

Competition

As is the case with many product offerings incorporating GPUs, NVIDIA is both a partner and a competitor. NVIDIA offers its own dense GPU servers for the computationally intensive task of training neural networks: the DGX-1 (8 GPUs) and DGX-2 (16 GPUs, configured into a memory fabric using the NVswitch technology). It also offers reference architectures based on these for its system OEM partners using the HGX brand, and those OEM partners could eventually become competitors to Watson Machine Learning Accelerator.

Intel is another hardware provider looking to capitalize on the growth of deep-learning workloads. Its long-awaited Intel Nervana Neural Network Processor (NNP-I) was officially announced at CES 2019. Intel also offers a variety of competing software tools for optimizing deep-learning workloads including an open source library for ML kernel optimization (called MKL DNN) and a distributed deep-learning framework for Apache Spark, called BigDL.

Finally, the growing AI market has attracted the attention of a variety of upstarts. As new accelerator offerings emerge from the close to 40 startups working on specialist chip architectures for AI, new systems offerings are likely to emerge. Frontrunners include Cerebras, Graphcore, Horizon Robotics and Wave Computing. The companies that build systems around those chips would be future competition for Watson Machine Learning Accelerator, although obviously, IBM will have been in the market for some time by then.

SWOT Analysis

STRENGTHS

Watson Machine Learning Accelerator provides an extensive number of tools that address several of the common concerns of data scientists. By bringing together IBM technology and expertise across the stack, IBM has created a software offering that should accelerate enterprise machine learning.

WEAKNESSES

The number of features contained within IBM Watson Accelerator is both impressive and overwhelming. Although the product targets more technical users, IBM could benefit from simplifying the toolset for end users.

OPPORTUNITIES

Although Watson Machine Learning Accelerator is targeted at the higher-end portion of IBM's compute portfolio right now, we're in the early stages of what is likely to be a wider market opportunity for machine-learning accelerators in the future.

THREATS

The majority, if not all, of the players in this space have competing acceleration offerings, and for many customers, it is the hardware – not the abstraction software – that will most influence purchasing decisions.