AI Deep Learning with IBM Spectrum Discover, Spectrum Scale and Cloud Object Storage

Silverton Consulting, Inc. StorInt™ Briefing

Introduction

Two main advances have brought artificial intelligence (AI), machine learning (ML) and most recently, deep learning (DL) out of the labs and into the enterprise spotlight: GPUs and large data collections. The data amassed by large companies in general and hyper-scalers in particular is growing by petabytes (PB) to exabytes (EB) a year. Combining this ever-increasing mass of data with GPU computational resources (thousands of GPU cores/card) is ideal for supplying DL functionality.

AI DL is a data-intensive process, so it's not surprising that IBM® Storage solutions can play an integral part in DL training and deployment. Some IBM Storage solutions that are especially useful for AI DL workflows include:

- **IBM Spectrum Discover**, a modern metadata management solution that is an excellent choice to help identify, classify, and label unstructured data across heterogeneous storage on-premises and in the cloud, as well as curate specific data sets for AI DL workflows.
- **IBM Spectrum Scale**, a flexible, high-throughput, unstructured data solution that is especially useful for ingesting massive amounts of data needed for AI DL; iterating over data through DL model architecture, design and training; and performing any data processing needed for inferencing and adaptation during model deployment.
- **IBM Cloud Object Storage (IBM COS)**, a hyper-scale, geographically dispersed object store solution useful for storing large amounts of data for long term analysis, archiving DL training data as well as AI deployment and adaptation data used for model development, use and compliance.

AI workflows

AI has come a long way since the middle of the last century, evolving from rules-based logic to expert systems to fuzzy logic to ML and now to DL as a form of ML. DL has become the dominant form of AI ML today because it makes better use of data and can more easily identify features of interest than other ML techniques. With DL today, almost any organization from 1 to 100,000 employees can make effective use of AI.

Data is key

The typical AI DL project begins and ends with data. While organizations have increasing stores of data, finding the right data to train DL models requires skill and care. Thus, the first step in an AI DL workflow is to find data suitable for model training, which requires identifying and separating out training data from an organization's vast stores of unstructured data.

Most DL training uses files or objects that contain media, text, data feeds, etc., and is best done on lots of small datasets. But just having data is not enough. It's important to select the right data, fill in any missing information and eliminate any dross to create proper training, validation and test sets. This process, called data and feature preparation or engineering, is essential to proper AI DL.

Training AI deep learning models

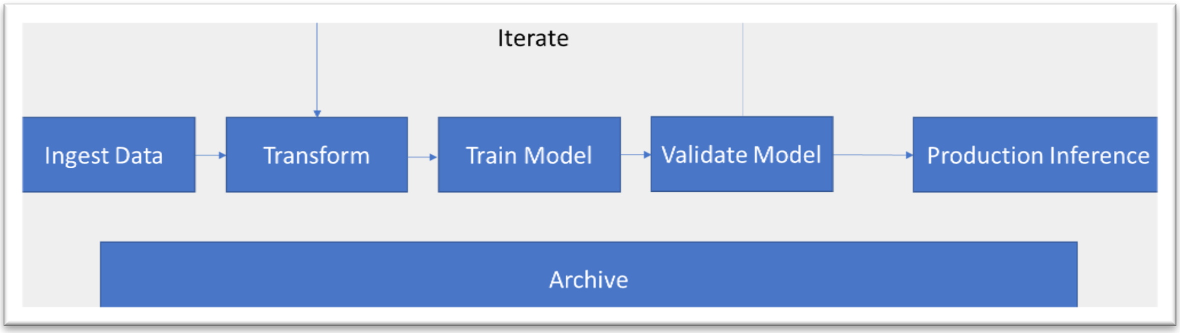
In the next step, all that data is fed into an DL neural network model instance for training, validation and testing. Training data generally takes the form of a feature set and a human- or machine- validated prediction or classification. For example, in real estate models used to predict median house pricing, the feature set may include median income of an area, average number of bedrooms, recent sale prices, etc. and a resulting media house price for the area. Hundreds to millions of these feature sets/predictions can then be used to train an AI DL model.

DL model training requires both **human engineering** and **automated computation**. Human engineering is needed to architect and design a DL neural network model, which involves establishing model hyper-parameters such as number of neural network layers, number of nodes per layer, accuracy/loss metrics (statistical measures of neural network prediction or classification correctness), etc.

Hyper-parameter selection typically occurs by selecting a sample set of model hyper-parameters, performing a complete pass against training data and calculating the current model's accuracy/loss metric based on the validation data (input parameter data and the [validated] correct prediction/classification for that data). Model hyper-parameters are then adjusted manually, after which the whole training process needs to be rerun. Hyper-parameter setting continues until the model design generates suitable accuracy over the training-validation data.

Automated computation takes place during each data pass and is used to adjust the model neural network layer weights in order to achieve better accuracy and lower loss levels. Many GPU cores are especially useful here. As GPU cards are expensive, having data presented quickly through caching, tiering and low-latency storage is essential to keeping GPU cores busy and not having to wait around for data to be made available.

The final step uses the training and validation data to train the DL model, which is qualified against the testing data. If suitable accuracy results, e.g. median housing price predictions for a region are within \$2.5K or a % of actual values, the AI DL model can be deployed. If insufficient accuracy results, new training data is needed, and the whole process must start over.



Using AI deep learning models

AI DL models operate by accessing data to make inferences, predictions or classifications. DL models in use typically run in two phases: 1) a **deployment phase**, where the model is used to perform some inferencing prediction on incoming data, and 2) an **adaptation phase**, where new predictions, input data and levels of accuracy are used to adapt or fine-tune the model for a specific use case.

For example, once a DL model has been trained to recognize general speech it can be deployed in speech recognition. In actual operation, humans may correct mistakes. This correction data can be used to adapt the speech recognition DL model for a specific person or environment. If these corrections are more general in nature, they can then be used during a new DL training pass.

Moreover, DL model deployment can take many forms. For instance, DL models are deployed at the edge in self-driving vehicles through real-time access to vehicle sensors. In fraud detection models, deployment occurs in data centers, processing card transaction feeds in near real time. In disease diagnosis models, deployment may occur in a hospital/clinic mini-data center, processing diagnostic imaging, blood assays and other clinical information in a batch process.

Furthermore, DL model deployment data is used in the adaptation phase to fine-tune a model instance for a specific use case. For example, a fraud detection model can be adapted to ignore designated transactions for a particular person. For adaptation that applies beyond a specific use case, adaptation data and a validated prediction/classification are added to the model's training set, requiring a new DL model training pass.

DL model training is never complete. Thus, once in deployment, new and old training data need to be archived for use during subsequent training passes. In addition, as some DL models deal in safety, financial or other critical domains, archived model training, adaptation and deployment data help support legal challenges and compliance regimens.

IBM Storage solutions

As discussed earlier, IBM has a number of storage solutions that can play a vital role during an organization's AI DL model training and deployment.

IBM Spectrum Discover

IBM Spectrum Discover is a new metadata management solution that enterprises can use to help manage PB to EB of unstructured data stores. For AI purposes, it essentially provides visibility into file and object metadata needed to identify, understand and select data for DL training.

IBM Spectrum Discover can map out storage consumption and provide searchable metadata indexes for unstructured data. Such indices can be especially useful for DL data selection. Spectrum Discover connects to both IBM and third-party file and object storage systems both on-premises and in the cloud to unify metadata from unstructured data, wherever it resides. Spectrum Discover supports deep inspection of many file types to extract metadata from file headers and use it to classify and label data. As such, Spectrum Discover can automatically identify as well as label data containing certain types of Personally Identifiable Information (PII). It also provides an extensible platform for more complex solutions via its Action Agent API. Organizations can use Action Agents to extend the solution's capabilities beyond indexing data, to identify internal field layouts of proprietary data formats, or to perform DL data discovery and feature engineering.

IBM Spectrum Discover comes as a software-only solution that installs as a VMware virtual appliance on any qualified server. As it provides essential support for DL training data discovery and selection, it will most likely reside in data centers or running on servers residing in the cloud.

IBM Spectrum Scale

IBM Spectrum Scale is a flexible, highly parallelized, high-throughput, exa-scale solution for unstructured data. It can make use of multiple backend storage systems and media and support all major file and object storage protocols. Select IBM Spectrum Scale functionality useful for AI DL workflows and deployment include:

- **High-throughput, parallelized, rapid data access**: Spectrum Scale has long been a favorite of high-performance computing (HPC) environments that require extremely large amounts of data throughput across thousands of servers and hundreds of storage nodes. It is ideal for capturing and presenting massive data streams used during DL model training and keeping up with near-real-time data flow during model deployment-adaptation.
- **Flexible storage support**: Spectrum Scale can support any storage media, including tape, disk and SSDs, as well as the cloud. It provides automated storage tiering to keep GPU compute cores from going idle during training and can archive all data encountered during deployment.

- **Countless files/objects support**: Spectrum Scale supports millions to billions of small to large files/objects in a hierarchical directory name space useful for DL data discovery, training and deployment archiving.

IBM Spectrum Scale can do just about everything needed to support enterprise AI DL data processing requirements, including ingesting massive amounts of data for training, iterating model training over vast quantities of training data, supporting data flows for DL model deployment-adaptation, and archiving all that data for future rounds of training and compliance.

Spectrum Scale comes as a software-only solution, server-bundled solution or cloud service. Server-bundled options offer quicker deployment, but customers may elect to deploy Spectrum Scale on their own storage or take advantage of its many cloud service options that are available.

IBM Cloud Object Storage

IBM COS supports a massive object storage repository that can span TB to PB or even EB of data. Indeed, IBM COS is used in the IBM Cloud™, as its own object storage backend and is integrated into the IBM Analytics portfolio and most cloud services including databases and SQL search tools. As more and more AI and Analytics tools leverage the S3 REST API such as the S3A interface for Spark and Hadoop or are written directly using S3 API, object storage will continue to grow in use. The compatibility of cloud native applications using the S3 API make it a great choice for hybrid cloud deployment.

Object storage is a relatively new form of data. It can be immutable and even locked down for compliance data, it exists in a flat name space (buckets), it has customizable metadata, and is accessed via RESTful S3 interfaces. Given the massive amounts of data used in DL training and deployment, object storage can be ideal for use as a repository for DL data.

IBM COS functionality especially relevant to AI DL workflows and deployment-adaptation includes:

- **Large, economical object store**: IBM COS supports TB to EB-sized object stores, uses low-cost storage media and supports billions of objects. Data is easily accessible from any location. Data reliability and security are designed in to the architecture. As such, it is useful for AI DL data storage and long-term archiving.
- **Geographic dispersion of data**: IBM COS supports geographical dispersion of data across vast distances without resorting to replication. It is extremely convenient for training data and model deployment data residing in multiple locations throughout the world. Data can be concurrently accessed from any location and from multiple access points proving high availability and throughput.

- **Immutable data store**: IBM COS supports immutable data that can be locked down for a set duration of time, which can support compliance requirements for DL data.

IBM COS is available as a software-only solution, as an appliance and in the IBM Cloud. As such, it can be deployed anywhere large amounts of AI DL data happens to reside.

Summary

Data is critical to today's AI DL workflows, including model training, deployment and adaptation. Although organizations amass lots of data, finding, selecting and transforming data suitable for AI DL training is difficult. Data discovery for DL training requires visibility into metadata and other constituent parts of an organization's unstructured data. Here, IBM Spectrum Discover can be especially useful.

Once appropriate DL training data has been found, a multi-phase, iterative process begins, which incorporates numerous design parameter adjustments to create an optimal DL training model. Once DL model design is complete, training begins with yet another pass over the training and testing data. IBM Spectrum Scale supports DL model training by ingesting and serving up all those datasets in a timely and efficient manner.

As more data comes in during DL model deployment and adaptation, the DL model can quickly process it to make inferences, classifications and predictions. IBM Spectrum Scale can be used to keep up with real-time and near-real-time data flows during deployment and adaptation.

Finally, all that training, deployment and adaptation data needs to be archived for future training and compliance purposes. IBM COS offers the flexibility and storage economics needed to archive the massive amounts of data used during DL model training and use.

In sum, IBM offers the storage technologies needed to enable organizations to implement AI DL for just about any purpose imaginable. For more information on IBM storage solutions for AI and Big Data: https://www.ibm.com/it-infrastructure/storage/ai-infrastructure.

Silverton Consulting, Inc., is a U.S.-based Storage, Strategy & Systems consulting firm offering products and services to the data storage community.

Disclaimer: This document was developed with International Business Machines Corporation (IBM) funding. Although the document uses

publicly available material from various sources, including IBM, it does not necessarily reflect the positions of such sources on the issues addressed.