Problem: Data Scientist Productivity

Waiting on resources
Experimenting with models takes time
Experimenting with hyper parameters takes time
Waiting on execution
Iterating over and over to improve accuracy

Also... Domain Scientist Productivity – enabling users that don’t have deep data science skills
Basic Workflow for ML/DL Training & Execution

Data Source
- Traditional Business Data
- Sensor Data
- Data from collaboration partners
- Data from mobile apps & social media
- Legacy Data

Data Preparation
- Data Ingestion
- Pre-Processing
- Training Dataset
- Testing Dataset

Model Training
- AI Deep Learning Frameworks (TensorFlow, PyTorch, ...)
- Distributed & Elastic Training for Deep Learning
- Parallel Hyper-Parameter Search & Optimization
- Monitor & Advise

Inference
- Deploy in Production using Trained Model (Rest API)
- Trained Model Life Cycle Management

Multi-Tenant, Shared Services Architecture
example #1
Auto Hyper-Parameter Tuning
Building a model – and **what’s a hyperparameter**

Each layer has a series of **weights** and **biases**

These are modified as the model is trained to produce higher accuracy

These weights and biases are held in matrices

The amount of change each iteration is called the **learning rate**
Building a model – and **what’s a hyperparameter**

Deep Learning training requires the system to learn from the data.

This requires multiple **iterations** or **epochs**

Over each iteration or epoch the system looks at all of the training data.

The model is refined based on the **loss** from previous run.
Auto Hyper-Parameter Tuning with WML Accelerator

- Data scientists run 100s of jobs with different Hyper-parameters
  - Learning rate, Decay rate, Batch size, Optimizers (GradientDecedent, Adadelta, Momentum, RMSProp, ..)
- Auto-Tuner searches for good hyper-parameters by launching 10s of jobs in parallel with 10,000s of iterations & selecting the best ones
  - 4 search algorithms: Random, Tree-based Parzen Estimator (TPE), Bayesian, Hyperband

![WML Accelerator Auto-Tuner (DL Insight)]

- Data
- DNN Model
- IBM WML Accelerator
- NVIDIA GPU-Accelerated IBM Power9 Servers

---

WML Accelerator Auto-Tuner (DL Insight)

- Job 1
- Job 2
- Job n
- Monitor & Prune
- Select Best Hyperparameters
Hyperparameter search & optimization (HPO)

• Find the best hyperparameters using cognitive algorithms running in parallel, refining the values as the search progresses.

• Supported Algorithms
  - Random Search
  - Bayesian
  - TPE
  - Hyperband
  - More to come…
Training Visualization
Monitor, Analyze, Optimize

Recommended Learning rate to re-train the model.

Suggestion
learning_rate=0.15
Two more examples

- IBM Bayesian Optimization Accelerator
- IBM Power AI Vision
What is Bayesian Optimization?

Bayesian optimization is a sequential design strategy for global optimization. **IBM BAO (‘19Q4)**

Many workflows require you to find a powerful set of parameters solve a problem. The challenge is finding those parameters robustly in as little time as possible.

**Applied to Computational Chemistry**

BOA accelerated workflow uses 1/3 of the calculations to achieve 4 orders of magnitude resolution increase

**Applied to Engineering Design**

BOA performed in 19 hours and ~30 simulations what an expert designer would do in 3 weeks

**Applied to Drug Discovery**

Brute force methods of screening require 20,000 experiments. BOA accelerated method required ~200
PowerAI Vision: "Point-and-Click" AI for images & video

Label Image or Video Data

Auto-Train AI Model
(no coding, just point and click)

Package & Deploy AI Model
IBM Machine Learning Accelerator - software

**AI for Data Scientists and non-Data Scientists**

- **Watson Machine Learning Community Edition**
- **WML Accelerator**

**Accelerated Infrastructure**

**Deep Learning Impact (DLI) Module**
- Data & Model Management, ETL, Visualize, Advise

**PowerAI Vision**
- Auto-DL for Images & Video
  - Label
  - Train
  - Deploy

**H2O Driverless AI**
- Auto-ML for Text & Numeric Data, NLP
  - Import
  - Experiment
  - Deploy

**WMLA CE : Open Source ML Frameworks**
- TensorFlow
- Caffe
- PyTorch
- Large Model Support (LMS)
- Distributed Deep Learning (up to 4 nodes)

**IBM Spectrum Conductor with Spark**
- Cluster Virtualization,
  Dynamic Resource Orchestration,
  Multiple Frameworks, Distributed Execution Engine

**WML Accelerator**
- Distributed Deep Learning (DDL – 1000s of nodes)
- Auto Hyper-parameter Tuning

**IBM Machine Learning Accelerator**
- software
  - Distributed Deep Learning (up to 4 nodes)

**Accelerated Infrastructure**
- IBM/NVIDIA Accelerated Servers
- Storage
IBM Mini-CORAL SUMMIT Starter Kit: What’s Included

Starter Configuration with (6) IBM Power Servers and AI/ML Software and Services
✓ IBM Machine Learning Accelerator software suite; RHEL 7.6
✓ 5 Days Lab Services to stand up: on-site AI/ML Workshop
✓ Software Installed in Manufacturing to accelerate time-to-value

Start your AI journey with a complete environment enabling data scientists to unlock advanced insights to drive business benefits.

Benefit from AI model learning times measured in minutes not hours

Use open source frameworks and leading-edge AI technology from IBM Research

Data Scientists Require End to End, Scalable HPC Enterprise Services to Optimize Pipeline Workflow
Thank You