How do ML parameters impact system performance?

A comparison of Spark MLib and Tensorflow

David Adeboye

Spark MLlib

Spark is a unified analytics engine for big data processing, excelling at iterative computation.

The main abstraction Spark provides is a resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel.

MLlib consists of common learning algorithms and utilities including classification, regression, clustering etc.

Focus: Distributed Computation

Tensorflow

Tensorflow is a (distributed) computational framework (primarily created) for building machine learning models.

Tensorflow is a framework to define and run computations involving tensors, which are a generalization of vectors and matrices to potentially higher dimensions.

Tensorflow natively supports distributed training over multiple processing units with minimal changes

---

Focus: **Dataflow and differentiable programming**

MNIST

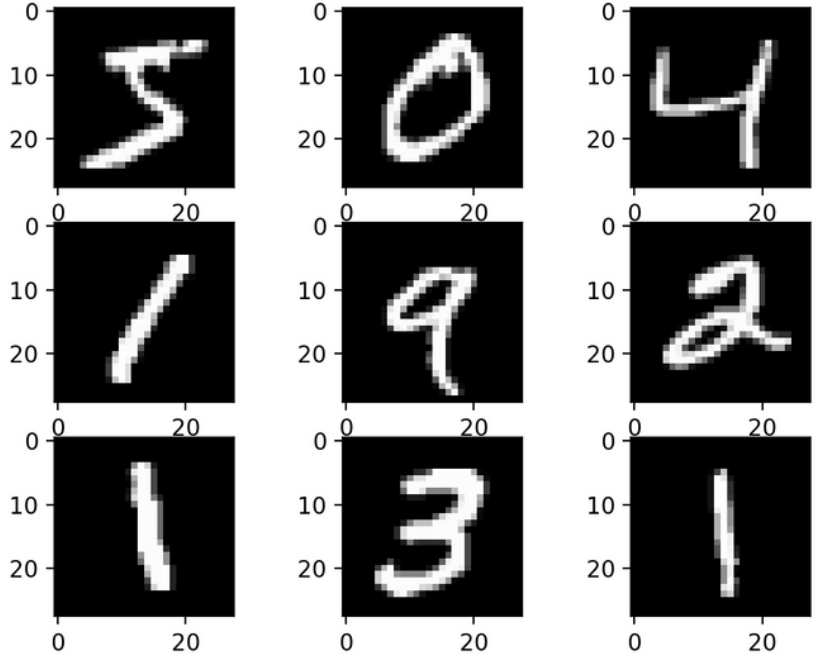
Database of handwritten digits

Training set of 60,000 examples, with a test set of 10,000 examples

Normalized to fit into a 28x28 pixel bounding box and anti-aliased

No pre-processing

2 multilayer perception classifier 784-800-10 (from MNIST Wikipedia page)



How do ML parameters impact system performance?

Project Goal

The project to build two convolution neural networks, one in Spark MLlib and one in TensorFlow and run them on a cluster and compare their relative performance.

Compare the qualitative differences between training Spark MLlib + Tensorflow models.

Can we trade-off accuracy for large gains in system resources?

Evaluation

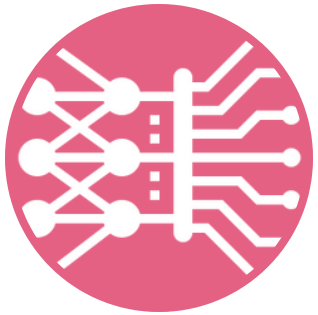
ML Parameters

Input Data Size

Hyperparameters

No. of iterations

Learning Rate



System Metrics

Training Time

Testing Time

Memory Usage

CPU Usage

Timeline

Research
Become aquatinted with Spark MLib and Tensorflow for CNNs

Build
Build and evaluate models for each framework

Prepare
Build cluster

Evaluate!

Extension
Explore TensorFlow on Spark

Extension: TensorFlowOnSpark

Library by Yahoo to enables distributed deep learning on a cluster of computers

Integrate TensorFlow and Spark pipeline, combing the advantages of both frameworks

Support all TensorFlow functionalities: synchronous/asynchronous training, model/data parallelism, inferencing and TensorBoard.

Allow datasets on HDFS and other sources pushed by Spark or pulled by TensorFlow.

Github: **yahoo/TensorFlowOnSpark**

Thanks for listening