DLI Accelerated Data Science Teaching Kit

Module 14 Lab:
Image Classification with RAPIDS-based Random Forest

OBJECTIVE

RAPIDS is a suite of open-source software libraries and APIs for executing data science pipelines entirely on GPUs—and can speed up training of machine learning models. Random forest is an ensemble model built with multiple tree models. Training random forest will cost a huge amount of time on big data. RAPIDS can be employed to speed up training of random forest. The goal of this lab is to speed up training of random forest on image classification with RAPIDS.

PREREQUISITES

Install main Python packages below.

- **numpy** is the fundamental package for scientific computing with Python.
- **pandas** is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool via Python.
- **sklearn** is contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- **cudf** is built based on the Apache Arrow columnar memory format, cuDF is a GPU DataFrame library for loading, joining, aggregating, filtering, and otherwise manipulating data.
- **cuml** is is a suite of libraries that implement machine learning algorithms and mathematical primitives functions that share compatible APIs with other RAPIDS projects.

INSTRUCTIONS

- Downloading *MNIST data set* employed for validating the model (*http://yann.lecun.com/exdb/mnist/)*.
- Splitting the *MNIST data set* into training and testing data
- Preparing training and testing data for training RAPIDS based random forest
- Training random forest classifier with sklearn on training data and recording the training time
- Evaluating the random forest classifier on testing data with accuracy
- Training RAPIDS based random forest classifier with sklearn on training data and recording the training time

- Evaluating the RAPIDS based random forest classifier on testing data with accuracy
- Comparing the accuracy between random forest classifier and RAPIDS based random forest classifier