CHAPTER 41

Google Cloud Machine Learning Engine (Cloud MLE)

The Google Cloud Machine Learning Engine, simply known as Cloud MLE, is a managed Google infrastructure for training and serving "large-scale" machine learning models. Cloud ML Engine is a part of GCP AI Platform. This managed infrastructure can train large-scale machine learning models built with TensorFlow, Keras, Scikit-learn, or XGBoost. It also provides modes of serving or consuming the trained models either as an online or batch prediction service. Using online prediction, the infrastructure scales in response to request throughout, while with the batch mode, Cloud MLE can provide inference for TBs of data.

Two important features of Cloud MLE is the ability to perform distribution training and automatic hyper-parameter tuning of your models while training. The big advantage of automatic hyper-parameter tuning is the ability to find the best set of parameters that minimize the model cost or loss function. This saves time of development hours in iterative experiments.

The Cloud MLE Train/Deploy Process

The high-level overview of the train/deploy process on Cloud MLE is depicted in Figure 41-1:

1. The data for training/inference is kept on GCS.
2. The execution script uses the application logic to train the model on Cloud MLE using the training data.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_41