Containerizing Deep Learning Workloads in Intel® Xeon® E3 Cluster for AI Web Applications

Srivignesh Pacham Sri Srinivasan, Intel Student Ambassador, Arizona State University

Abstract: Deep learning model based web applications are data intensive and involve complex model computation in each task. These models are built in various deep learning frameworks. There is high complexity involved in loading and computing the test results of the models. Hence, we propose to use containers to decouples the app environment from the running host and encapsulates all dependencies in a single portable unit.

Docker is an open container platform for developers to build, ship, and run distributed applications on data center VMs, or the cloud. Hashicorp Nomad is a tool for managing a cluster of machines and running applications on them. Nomad abstracts away machines and the location of applications, and instead enables users to declare what they want to run and Nomad handles where they should run and how to run them.

Intel® Caffe and Intel® Tensorflow is dedicated to improving performance of these deep learning frameworks when running on CPU, in particular Intel® Xeon processors and Intel® Xeon Phi processors. Model Zoo is an ongoing project to collect complete models, with python scripts, pre-trained weights as well as instructions on how to build and fine tune these models. This includes ImageNet models such as VGG-16, AlexNet, GoogLeNet, LeNet, and more.

Workflow Approach

- Nomad agent configures the Xeon CPU cluster and creates a job listening server.
- When a Nomad job is posted it creates a new allocation and schedules a Docker containers which has required dependencies such as Intel® Caffe and Intel® Tensorflow.
- Deep learning model is executed inside the container and the results are stored in the log for each allocation.
- Web Server API fetches the results from these logs and displayed them to the client.
- Docker containers are destroyed as soon as the request is processed.

Nomad Workflow

- Provision
- Schedule
- Terminate

Launch Docker Containers

Deep learning Model Zoo Workload Execution

- Provides ability to run multiple framework on a single running host without missing dependencies.
- Operationally simple and easy to scale the application.
- Improved performance because of Intel architecture optimizations.
- Easy to upgrade framework to latest frameworks.
- The Model compile, build and shipping cycle can be improved significantly.

Thanks to Intel Student Ambassador Forum for providing access to Xeon CPU Clusters, guidance and support for Intel® Caffe and Intel® Tensorflow.