

# The Nebula Benchmark Suite: Implications of Lightweight Neural Networks

Bogil Kim, Sungjae Lee, Chanhoo Park, Hyeonjin Kim, and William J. Song  
School of Electrical Engineering, Yonsei University

Jan. 2020 (Version 1.0)

## 1. Introduction

*Nebula* is a lightweight benchmark suite for neural networks. Recent neural networks become increasingly deeper and larger. This trend puts great challenges on the modeling and analysis of computer systems in that it takes longer execution time to process a large number of operations and sizable data. Nebula tackles this challenge by taking the opposite direction based on an observation that the computations of neural networks are mainly comprised of matrix and vector operations. Nebula benchmarks drastically reduce the number of operations for an expectation that the lightened networks will still have similar architectural behaviors as the full-fledged neural networks, and thus it is not always necessary to run the complete networks with sizable data if the purpose is to characterize micro-architectural behaviors in computer systems. Results based on hardware measurements prove that Nebula benchmarks indeed meet the similarity and affordability goal.

## 2. Prerequisite, Download, and Installation

The Nebula benchmark suite implements a C++ framework that facilitates the composition of diverse neural networks, similar to the well-known Caffe framework. The framework standardizes the implementation of network and layer classes for easier extension to new type of neural networks. The standardization makes it convenient to interface with them and share acceleration libraries among the various implementations of neural networks.

Nebula requires g++ compiler to build on CPU, nvcc on GPU, and OpenCV for image processing. It utilizes several acceleration libraries encompassing OpenBLAS for CPU and CUDA libraries (e.g., cuBLAS and cuDNN) to speed up the computations of neural networks, but the implementations are not limited to those external libraries. To install prerequisite to run the Nebula benchmark, use the following command in terminal.

```
[required] * g++  
[required] * OpenCV  
[optional] * nvcc  
[optional] * OpenBLAS for CPU  
[optional] * cuBLAS for GPU  
[optional] * cuDNN for GPU
```

It has been validated in Ubuntu 16.04 and 18.04. To obtain a copy of Nebula v1.0 (as of Jan., 2020) and to build it, use the following command in terminal.

```
$ git clone https://github.com/yonsei-icsl/nebula.git  
$ cd <working directory>/nebula  
$ ./nebula.sh build all
```

### 3. Running Nebula Framework

Nebula runs the neural network with network configuration file, dataset image, and weight file. Figure network configuration file is located in benchmark directory (benchmarks/vgg/network.cfg). You can obtain dataset image and weight file using the command *get\_data.sh* and *get\_weight.sh*, respectively.

```
$ ./get_data.sh
Which dataset? [ImageNet[I] / NIST[N] / MNIST[M]] I
Which size? [Large[L] / Medium[M] / Small[S]] S
Get data ID of ImageNet_S
--2020-04-16 19:11:14-- https://docs.google.com/uc?export=download&confirm=2bpp&id=1zCTB82sCmBcf0tCapOa0cd7tAcPp9ADe
Resolving docs.google.com (docs.google.com)... 172.217.25.206, 2404:6800:4004:81a::200e
Connecting to docs.google.com (docs.google.com)|172.217.25.206|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily

...

Connecting to doc-0o-a0-docs.googleusercontent.com (doc-0o-a0-docs.googleusercontent.com)|172.217.161.33|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [application/x-tar]
Saving to: ImageNet_S.tar

ImageNet_S.tar          [          <=>          ]  2.82G  5.15MB/s
```

```
$ ./get_weight.sh

Get file ID of vgg_S
/home/bogil/Publication/nebula/benchmarks/vgg
--2020-04-16 19:23:32-- https://docs.google.com/uc?export=download&confirm=&id=1L6GzG0Je43jd6sVWICFC8oCvEtP507ee
Resolving docs.google.com (docs.google.com)... 172.217.25.206, 2404:6800:4004:81a::200e
Connecting to docs.google.com (docs.google.com)|172.217.25.206|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily

...

Connecting to doc-00-2g-docs.googleusercontent.com (doc-00-2g-docs.googleusercontent.com)|172.217.161.33|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [application/octet-stream]
Saving to: input.wgh

input.wgh              [          <=>          ]  11.74M  5.31MB/s
```

With input files (i.e., network configuration file, dataset image, and weight file of network), you can run variable-sized neural networks. After executing the network, Nebula prints out the result of execution depending on the execution model (i.e., inference and training). The result of inference includes instantaneous accuracy, cumulative accuracy, and execution time. On the other hands, an instantaneous loss, and a cumulative loss, and runtime are printed out at the end of every iteration of training phase.

```
$ ./nebula.sh test vgg
```

