DALI

DU-09049-001 _v0.20.0 Beta Release | March 2020

Installation Guide

TABLE OF CONTENTS

Chapter 1. OVERVIEW

Deep learning applications require complex, multi-stage pre-processing data pipelines. Such data pipelines involve compute-intensive operations that are carried out on the CPU. For example, tasks such as: load data from disk, decode, crop, random resize, color and spatial augmentations and format conversions, are mainly carried out on the CPUs, limiting the performance and scalability of training and inference.

In addition, the deep learning frameworks have multiple data pre-processing implementations, resulting in challenges such as portability of training and inference workflows, and code maintainability.

NVIDIA® Data Loading Library™ (DALI) is a collection of highly optimized building blocks, and an execution engine, to accelerate the pre-processing of the input data for deep learning applications. DALI provides both the performance and the flexibility of accelerating different data pipelines as a single library. This single library can then be easily integrated into different deep learning training and inference applications.

Highlights of DALI are:

- Full data pipeline—accelerated from reading disk to getting ready for training/ inference.
- Flexibility through configurable graphs and custom operators.
- Support for image classification and segmentation workloads.
- Ease of integration through direct framework plugins and open source bindings.
- Portable training workflows with multiple input formats - JPEG, PNG (fallback to CPU), TIFF (fallback to CPU), BMP (fallback to CPU), raw formats, LMDB, RecordIO, TFRecord.
- Extensible for user specific needs through open source license.

Chapter 2.
DALI AND NVIDIA GPU CLOUD

DALI is pre-installed in the NVIDIA® GPU Cloud™ (NGC) TensorFlow, PyTorch, and MXNet containers in version 18.07 and later.

Chapter 3. INSTALLING DALI

DALI can be installed either directly using a pre-built binary or by compiling the sources from GitHub.

3.1. Installing Prebuilt DALI Packages

3.1.1. Prerequisites

Ensure you meet the following minimum requirements:

- Linux x64
- NVIDIA Driver (384.xx or later driver releases) supporting CUDA 9.0 or later
- One or more of the following deep learning frameworks:
  - MXNet 1.3 or later
    - Version 1.3 from the Python package with the following command:

      ```
      pip install mxnet-cu90==1.3.0
      ```
  - PyTorch 0.4
  - TensorFlow 1.7 or later.

3.1.2. Binary Installation

- To install the CUDA 9.0-based DALI build using **pip**, execute:

  ```
  $ pip install --extra-index-url https://developer.download.nvidia.com/
  compute/redist/cuda/9.0 nvidia-dali
  ```
- To install the CUDA 10-based DALI build using **pip**, execute:

```
$ pip install --extra-index-url https://developer.download.nvidia.com/
compute/redist/cuda/10.0 nvidia-dali
```

Starting with DALI 0.6.1 the **nvidia-dali** package no longer contains prebuilt versions of the DALI TensorFlow plugin, so you need to install the DALI TensorFlow plugin for the currently installed version of TensorFlow.

- To install the DALI TensorFlow plugin for the CUDA 9.0-based DALI build, execute:

```
pip install --extra-index-url https://developer.download.nvidia.com/compute/
redist/cuda/9.0 nvidia-dali-tf-plugin
```

- To install the DALI TensorFlow plugin for the CUDA 10-based DALI build, execute:

```
pip install --extra-index-url https://developer.download.nvidia.com/compute/
redist/cuda/10.0 nvidia-dali-tf-plugin
```

- Installing the **nvidia-dali-tf-plugin** package will install **nvidia-dali** and its dependencies if these dependencies are not already installed.
- The package **tensorflow-gpu** must be installed before attempting to install **nvidia-dali-tf-plugin**.
- The package **nvidia-dali-tf-plugin** strictly requires that **nvidia-dali** be of the exact corresponding version. Thus, installing the latest version of **nvidia-dali-tf-plugin** will replace any older **nvidia-dali** versions that are already installed, with the latest version of **nvidia-dali**. To work with older versions of DALI, provide the version explicitly to the pip install command, as below:

```
OLDER_VERSION=0.6.1
pip install --extra-index-url https://developer.download.nvidia.com/
compute/redist nvidia-dali-tf-plugin==$OLDER_VERSION
```

3.2. Compiling DALI From Source (Bare metal)

For the most up-to-date compilation instructions, see Compiling DALI from source.

Chapter 4. EXECUTING INPUT PIPELINE WITH FULL TRAINING

After you've installed DALI, you can run pre-configured models accelerated by DALI, on MXNet, TensorFlow, PaddlePaddle and PyTorch frameworks. Each of the following samples offload image loading and augmentation operations onto GPUs.

For more information, see Use Cases.

Chapter 5. FFMPEG

This software uses code of FFmpeg licensed under the LGPLv2.1 and its source can be downloaded here.

FFmpeg was compiled using the following command line:

```
./configure \
 --prefix=/usr/local \
 --disable-static \
 --disable-all \
 --disable-autodetect \
 --disable-iconv \
 --enable-shared \
 --enable-avformat \
 --enable-avcodec \
 --enable-avfilter \
 --enable-protocol=file \
 --enable-demuxer=mov,matroska \
 --enable-bsf=h264_mp4toannexb,hevc_mp4toannexb
./make
```

Chapter 6. UNINSTALLING DALI

Uninstall DALI.

```
pip uninstall -y nvidia-dali
```

Notice

THE INFORMATION IN THIS GUIDE AND ALL OTHER INFORMATION CONTAINED IN NVIDIA DOCUMENTATION REFERENCED IN THIS GUIDE IS PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE INFORMATION FOR THE PRODUCT, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the product described in this guide shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

THE NVIDIA PRODUCT DESCRIBED IN THIS GUIDE IS NOT FAULT TOLERANT AND IS NOT DESIGNED, MANUFACTURED OR INTENDED FOR USE IN CONNECTION WITH THE DESIGN, CONSTRUCTION, MAINTENANCE, AND/OR OPERATION OF ANY SYSTEM WHERE THE USE OR A FAILURE OF SUCH SYSTEM COULD RESULT IN A SITUATION THAT THREATENS THE SAFETY OF HUMAN LIFE OR SEVERE PHYSICAL HARM OR PROPERTY DAMAGE (INCLUDING, FOR EXAMPLE, USE IN CONNECTION WITH ANY NUCLEAR, AVIONICS, LIFE SUPPORT OR OTHER LIFE CRITICAL APPLICATION). NVIDIA EXPRESSLY DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY OF FITNESS FOR SUCH HIGH RISK USES. NVIDIA SHALL NOT BE LIABLE TO CUSTOMER OR ANY THIRD PARTY, IN WHOLE OR IN PART, FOR ANY CLAIMS OR DAMAGES ARISING FROM SUCH HIGH RISK USES.

NVIDIA makes no representation or warranty that the product described in this guide will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this guide. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this guide, or (ii) customer product designs.

Other than the right for customer to use the information in this guide with the product, no other license, either expressed or implied, is hereby granted by NVIDIA under this guide. Reproduction of information in this guide is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

Trademarks

NVIDIA, the NVIDIA logo, and Ampere, cuBLAS, CUDA, CUDA Toolkit, cuDNN, DALI, DIGITS, DGX, DGX-1, DGX-2, DGX Station, DLProf, GPU, Jetson, Kepler, Maxwell, NCCL, Nsight Compute, Nsight Systems, NvCaffe, NVLink, NVIDIA Deep Learning SDK. NVIDIA Developer Program, NVIDIA GPU Cloud, NVSHMEM, PerfWorks, Pascal, SDK Manager, Tegra, TensorRT, TensorRT Inference Server, Tesla, TF-TRT, Triton Inference Server, Turing, and Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

ARM, AMBA and ARM Powered are registered trademarks of ARM Limited. Cortex, MPCore and Mali are trademarks of ARM Limited. All other brands or product names are the property of their respective holders. “ARM” is used to represent ARM Holdings plc; its operating company ARM Limited; and the regional subsidiaries ARM Inc.; ARM KK; ARM Korea Limited.; ARM Taiwan Limited; ARM France SAS; ARM Consulting (Shanghai) Co. Ltd.; ARM Germany GmbH; ARM Embedded Technologies Pvt. Ltd.; ARM Norway, AS and ARM Sweden AB