# TABLE OF CONTENTS

Chapter 1. Overview Of TensorFlow ................................................................. 1
   1.1. Contents Of The NVIDIA TensorFlow Container ................................. 1

Chapter 2. Pulling The TensorFlow Container ............................................... 3

Chapter 3. Running A TensorFlow Container ................................................. 4

Chapter 4. Verifying TensorFlow .................................................................. 6

Chapter 5. Customizing And Extending TensorFlow ....................................... 7
   5.1. Benefits And Limitations To Customizing TensorFlow ......................... 8
   5.2. Example 1: Customizing TensorFlow Using Dockerfile ....................... 8
   5.3. Example 2: Customizing TensorFlow Using docker commit .................. 10

Chapter 6. TensorFlow Parameters ............................................................... 13
   6.1. Added And Modified Parameters .................................................... 13

Chapter 7. TensorFlow Environment Variables .......................................... 15
   7.1. Added Or Modified Variables ....................................................... 15

Chapter 8. Performance .............................................................................. 19
   8.1. Tensor Core Math ........................................................................ 19
       8.1.1. Float16 Training .................................................................... 19

Chapter 9. Troubleshooting ....................................................................... 21
   9.1. Support .................................................................................... 21
Chapter 1.
OVERVIEW OF TENSORFLOW

TensorFlow™ is an open-source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) that flow between them. This flexible architecture lets you deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device without rewriting code.

TensorFlow was originally developed by researchers and engineers working on the Google Brain team within Google’s Machine Intelligence research organization for the purposes of conducting machine learning and deep neural networks (DNNs) research. The system is general enough to be applicable in a wide variety of other domains, as well.

For visualizing TensorFlow results, the Docker® image also contains TensorBoard. TensorBoard is a suite of visualization tools. For example, you can view the training histories as well as what the model looks like.

For information about the optimizations and changes that have been made to TensorFlow, see the Deep Learning Frameworks Release Notes.

1.1. Contents Of The NVIDIA TensorFlow Container

This image contains source and binaries for TensorFlow. The pre-built and installed version of TensorFlow is located in the /usr/local/[bin,lib] directories. The complete source code is located in /opt/tensorflow.

To achieve optimum TensorFlow performance, there are sample scripts within the container image. For more information, see Performance.

TensorFlow includes TensorBoard, a data visualization toolkit developed by Google. Additionally, this container image also includes several built-in TensorFlow examples that you can run using commands like the following. These examples perform training of convolutional neural networks (CNNs). For more information, see MNIST For ML Beginners. The following Python commands run two of these examples:

```python
python -m tensorflow.models.image.mnist.convolutional
```
python -m tensorflow.models.image.cifar10.cifar10_multi_gpu_train

The first command uses the MNIST data set, for example, THE MNIST DATABASE. The second command uses the CIFAR-10 dataset, for example, The CIFAR-10 dataset.
You can pull (download) an NVIDIA container that is already built, tested, tuned, and ready to run. Each NVIDIA deep learning container includes the code required to build the framework so that you can make changes to the internals. The containers do not contain sample data-sets or sample model definitions unless they are included with the source for the framework.

Currently, you can access NVIDIA GPU accelerated containers in one of two ways depending upon where you doing your training. If you own a DGX-1™ or a DGX Station™, then you should use the NVIDIA® DGX™ container registry located at https://compute.nvidia.com. You can pull the containers from there and you can also push containers there into your own account on the nvidia-docker repository, nvcr.io.

If you are accessing the NVIDIA containers from a Cloud Server Provider such as Amazon Web Services (AWS), then you should first create an account at the NVIDIA NGC Cloud Services located at https://ngc.nvidia.com. After you create an account, the commands to use containers are the same for the DGX-1 and the DGX Station. However, currently, you cannot save any containers to the NVIDIA® GPU Cloud™ (NGC) container registry, nvcr.io if you are using NVIDIA® GPU Cloud™ (NGC). Instead you have to save the containers to your own Docker repository.

The containers are exactly the same, whether you pull them from the NVIDIA DGX container registry or the NGC container registry.

Before you can pull a container you must have Docker and nvidia-docker installed as explained in Preparing to use NVIDIA Containers Getting Started Guide. You must also have access and logged into the NGC container registry as explained in NGC Getting Started Guide.
Chapter 3.
RUNNING A TENSORFLOW CONTAINER

To run a container, you must issue the `nvidia-docker run` command, specifying the registry, repository, and tags. For example:

```
$ nvidia-docker run nvcr.io/nvidia/tensorflow:17.05
```

Before you can run an nvidia-docker deep learning framework container, you must have `nvidia-docker` installed. For more information, see the Preparing to use NVIDIA Containers Getting Started Guide.

1. Copy the command for the applicable release of the container that you want. The following command assumes you want to pull the latest container.

```
docker pull nvcr.io/nvidia/tensorflow:17.10
```

2. Open a command prompt and paste the pull command. The pulling of the container image begins. Ensure the pull completes successfully before proceeding to the next step.

3. Run the container image. A typical command to launch the container is:

```
nvidia-docker run -it --rm -v local_dir:container_dir nvcr.io/nvidia/tensorflow:<xx.xx>
```

Where:

- `-it` means interactive
- `--rm` means delete the image when finished
- `-v` means mount directory
- `local_dir` is the directory or file from your host system (absolute path) that you want to access from inside your container. For example, the `local_dir` in the following path is `/home/jsmith/data/mnist`.

```
-v /home/jsmith/data/mnist:/data/mnist
```

If you are inside the container, for example, `ls /data/mnist`, you will see the same files as if you issued the `ls /home/jsmith/data/mnist` command from outside the container.
**container dir** is the target directory when you are inside your container. For example, `/data/mnist` is the target directory in the example:

```
-v /home/jsmith/data/mnist:/data/mnist
```

**<xx.xx>** is the tag. For example, 17.06.

TensorFlow™ is run by importing it as a Python module:

```
$ python
>>> import tensorflow as tf
>>> hello = tf.constant('Hello, TensorFlow!')
>>> sess = tf.Session()
>>> sess.run(hello)
Hello, TensorFlow!
```

```
>>> a = tf.constant(10)
>>> b = tf.constant(32)
>>> sess.run(a+b)
42
```

You might want to pull in data and model descriptions from locations outside the container for use by TensorFlow or save results to locations outside the container. To accomplish this, the easiest method is to mount one or more host directories as Docker® data volumes.

In order to share data between ranks, NVIDIA® Collective Communications Library™ (NCCL) may require shared system memory for IPC and pinned (page-locked) system memory resources. The operating system’s limits on these resources may need to be increased accordingly. Refer to your system’s documentation for details.

In particular, Docker containers default to limited shared and pinned memory resources. When using NCCL inside a container, it is recommended that you increase these resources by issuing:

```
--shm-size=1g --ulimit memlock=-1
```

in the command line to

```
nvidia-docker run
```

For detailed instructions about the run command, see the NVIDIA Containers For Deep Learning Frameworks User Guide.
The simplest way to verify that TensorFlow is running correctly, is to run the examples that are included in the /nvidia-examples/ directory. Each example contains a README that describes the basic usage.
Chapter 5.
CUSTOMIZING AND EXTENDING TENSORFLOW

The nvidia-docker images come prepackaged, tuned, and ready to run; however, you may want to build a new image from scratch or augment an existing image with custom code, libraries, data, or settings for your corporate infrastructure. This section will guide you through exercises that will highlight how to create a container from scratch, customize a container, extend a deep learning framework to add features, develop some code using that extended framework from the developer environment, then package that code as a versioned release.

Currently, only the DGX-1 and DGX Station can push or store containers to the nvidia-docker repository, `nvcr.io`. The NVIDIA NGC Cloud Services can only store containers in a private repository outside of NVIDIA.

By default, you do not need to build a container. The NGC container registry NVIDIA container repository, `nvcr.io`, has a number of containers that can be used immediately including containers for deep learning as well as containers with just the CUDA® Toolkit™.

One of the great things about containers is that they can be used as starting points for creating new containers. This can be referred to as customizing or extending a container. You can create a container completely from scratch, however, since these containers are likely to run on GPUs, it is recommended that you at least start with a `nvcr.io` container that contains the OS and CUDA®. However, you are not limited to this and can create a container that runs on the CPUs which does not use the GPUs. In this case, you can start with a bare OS container from another location such as Docker. To make development easier, you can still start with a container with CUDA; it is just not used when the container is used.

The customized or extended containers can be saved to a user’s private container repository. They can also be shared with other users but this requires some administrator help.
It is important to note that all nvidia-docker deep learning framework images include the source to build the framework itself as well as all of the prerequisites.

**Attention** Do not install an NVIDIA driver into the Docker image at docker build time. The nvidia-docker is essentially a wrapper around docker that transparently provisions a container with the necessary components to execute code on the GPU.

A best-practice is to *avoid* docker commit usage for developing new docker images, and to use Dockerfiles instead. The Dockerfile method provides visibility and capability to efficiently version-control changes made during development of a Docker image. The Docker commit method is appropriate for short-lived, disposable images only.

For more information on writing a Docker file, see the best practices documentation.

### 5.1. Benefits And Limitations To Customizing TensorFlow

You can customize a container to fit your specific needs for numerous reasons; for example, you depend upon specific software that is not included in the container that NVIDIA provides. No matter your reasons, you can customize a container.

The container images do not contain sample data-sets or sample model definitions unless they are included with the framework source. Be sure to check the container for sample data-sets or models.

### 5.2. Example 1: Customizing TensorFlow Using Dockerfile

Before customizing the container, you should ensure the TensorFlow 17.04 container has been loaded into the NGC container registry using the `docker pull` command before proceeding. For example:

```bash
$ docker pull nvcr.io/nvidia/tensorflow:17.04
```

The Docker containers on nvcr.io also provide a sample Dockerfile that explains how to patch a framework and rebuild the Docker image. In the directory, `/workspace/docker-examples`, there are two sample Dockerfiles that you can use. The first one, `Dockerfile.addpackages`, can be used to add packages to the TensorFlow image. The second one, `Dockerfile.customtensorflow`, illustrates how to patch TensorFlow and rebuild the image.

```bash
FROM nvcr.io/nvidia/tensorflow:17.04

# Bring in changes from outside container to /tmp
# (assumes my-tensorflow-modifications.patch is in same directory as Dockerfile)
```
COPY my-tensorflow-modifications.patch /tmp

# Change working directory to TensorFlow source path
WORKDIR /opt/tensorflow

# Apply modifications
RUN patch -p1 < /tmp/my-tensorflow-modifications.patch

# Rebuild TensorFlow
RUN yes "" | ./configure && \
  bazel build -c opt --config=cuda \
  tensorflow/tools/pip_package:build_pip_package && \
  bazel-bin/tensorflow/tools/pip_package/build_pip_package /tmp/pip && \
  pip install --upgrade /tmp/pip/tensorflow-*.whl && \
  rm -rf /tmp/pip/tensorflow-*.whl && \
  bazel clean --expunge

# Apply modifications
WORKDIR /workspace

This DockerFile will rebuild the TensorFlow image in the same way as it was built in the original image. For more information, see Dockerfile reference.

To better understand the Dockerfile, let's walk through the major commands. The first line in the Dockerfile is the following:

FROM nvcr.io/nvidia/tensorflow:17.04

This line starts with the NVIDIA 17.04 version image for TensorFlow being used as the starting point.

The second line is the following:

COPY my-tensorflow-modifications.patch /tmp

It brings in changes from outside the container into your /tmp directory. This assumes that the my-tensorflow-modifications.patch file is in the same directory as Dockerfile.

The next important line in the file changes the working directory to the TensorFlow source path.

WORKDIR /opt/tensorflow

This is followed by the command to apply the modifications patch to the source.

RUN patch -p1 < /tmp/my-tensorflow-modifications.patch

After the patch is applied, the TensorFlow image can be rebuilt. This is done via the RUN command in the DockerFile/.

RUN yes "" | ./configure && \
  bazel build -c opt --config=cuda \
  tensorflow/tools/pip_package:build_pip_package && \
  bazel-bin/tensorflow/tools/pip_package/build_pip_package /tmp/pip && \
  pip install --upgrade /tmp/pip/tensorflow-*.whl && \
  rm -rf /tmp/pip/tensorflow-*.whl && \
  bazel clean --expunge
bazel clean --expunge

Finally, the last major line in the DockerFile resets the default working directory.

WORKDIR /workspace

5.3. Example 2: Customizing TensorFlow Using docker commit

This example uses the `docker commit` command to flush the current state of the container to a Docker image. This is not a recommended best practice, however, this is useful when you have a container running to which you have made changes and want to save them. In this example, we are using the `apt-get` tag to install a package that requires the user run as root.

- The TensorFlow image release 17.04 is used in the example instructions for illustrative purposes.
- Do not use the `--rm` flag when running the container. If you use the `--rm` flag when running the container your changes will be lost when exiting the container.

1. Pull the Docker container from the `nvcr.io` repository to your DGX™ system. For example, the following command will pull the TensorFlow container:

   ```bash
   $ docker pull nvcr.io/nvidia/tensorflow:17.04
   ```

2. Run the container on your DGX using the `nvidia-docker` command.

   ```bash
   $ nvidia-docker run -ti nvcr.io/nvidia/tensorflow:17.04
   ```

   ===============
   == TensorFlow ==
   ===============

   NVIDIA Release 17.04 (build 21630)

   Container image Copyright (c) 2017, NVIDIA CORPORATION. All rights reserved.
   Copyright 2017 The TensorFlow Authors. All rights reserved.

   Various files include modifications (c) NVIDIA CORPORATION. All rights reserved.
   NVIDIA modifications are covered by the license terms that apply to the underlying project or file.

   NOTE: The SHMEM allocation limit is set to the default of 64MB. This may be
3. You should now be the root user in the container (notice the prompt). You can use the command `apt` to pull down a package and put it in the container.

The NVIDIA containers are built using Ubuntu which uses the `apt-get` package manager. Check the container release notes Deep Learning Documentation for details on the specific container you are using.

In this example, we will install octave; the GNU clone of MATLAB, into the container.

```
# apt-get update
# apt install octave
```

You have to first issue `apt-get update` before you install Octave using `apt`.

4. Exit the workspace.

```
# exit
```

5. Display the list of running containers.

```
$ docker ps -a
```

As an example, here is some of the output from the `docker ps -a` command:

```
CONTAINER ID IMAGE CREATED
8db6076d82c4 nvcr.io/nvidia/tensorflow:17.04 3 minutes ago ...
```

6. Now you can create a new image from the container that is running where you have installed Octave. You can commit the container with the following command.

```
$ docker commit 8db6076d82c4 nvcr.io/nvidia_sas/tensorflow_octave:17.04
```

7. Display the list of images.

```
$ docker images
```

8. To verify, run the container again and see if Octave is actually there.

```
$ nvidia-docker run -ti nvvidia_sas/tensorflow_octave:17.04
================
== TensorFlow ==
================
```
NVIDIA Release 17.04 (build 21630)

Container image Copyright (c) 2017, NVIDIA CORPORATION. All rights reserved. Copyright 2017 The TensorFlow Authors. All rights reserved.

Various files include modifications (c) NVIDIA CORPORATION. All rights reserved. NVIDIA modifications are covered by the license terms that apply to the underlying project or file.

NOTE: The SHMEM allocation limit is set to the default of 64MB. This may be insufficient for TensorFlow. NVIDIA recommends the use of the following flags:
  nvidia-docker run --shm-size=1g --ulimit memlock=-1 --ulimit stack=67108864 ...

root@87e8dde4be6d:/workspace# octave
octave: X11 DISPLAY environment variable not set
octave: disabling GUI features
GNU Octave, version 4.0.0
Copyright (C) 2015 John W. Eaton and others.
This is free software; see the source code for copying conditions.
There is ABSOLUTELY NO WARRANTY; not even for MERCHANTABILITY or
FITNESS FOR A PARTICULAR PURPOSE. For details, type
'warranty'.

Octave was configured for "x86_64-pc-linux-gnu".

Additional information about Octave is available at http://www.octave.org.

Please contribute if you find this software useful.
For more information, visit http://www.octave.org/get-involved.html

Read http://www.octave.org/bugs.html to learn how to submit bug reports.
For information about changes from previous versions, type
'news'.

octave:1>

Since the Octave prompt displayed, Octave is installed.

9. If you are using a DGX-1 or DGX Station, and you want to save the container into your private repository (Docker uses the phrase "push"), then you can use the command docker push ....

$ docker push nvcr.io/nvidian_sas/tensorflow_octave:17.04

You cannot push the container to nvcr.io if you are using the NVIDIA NGC Cloud Services. However, you can push it to your own private repository. The new Docker image is now available for use. You can check your local Docker repository for it.
Chapter 6.
TENSORFLOW PARAMETERS

The TensorFlow container in the NGC container registry (nvcr.io) comes pre-configured as defined by the following parameters. These parameters are used to pre-compile GPUs, enable support for the Accelerated Linear Algebra (XLA) backend, and disable support for Google Cloud Platform (GCP) and the Hadoop Distributed File System (HDFS).

6.1. Added And Modified Parameters

In addition to the parameters within the Dockerfile that is included in the Google TensorFlow container, the following parameters have either been added for modified with the NVIDIA version of TensorFlow.

For parameters not mentioned in this guide, see the Google documentation.

6.1.1. TF_CUDA_COMPUTE_CAPABILITIES

The TF_CUDA_COMPUTE_CAPABILITIES parameter enables the code to be pre-compiled for specific GPU architectures.

The container comes built with the following setting, which targets Kepler, Maxwell, and Pascal GPUs:

```
TF_CUDA_COMPUTE_CAPABILITIES "5.2,6.0,6.1,7.0"
```

Where the numbers correspond to GPU architectures:

- 5.2 Maxwell
- 6.0+6.1 Pascal
- 7.0 Volta

6.1.2. TF_NEED_GCP
The TF_NEED_GCP parameter, as defined, disables support for the Google Cloud
Platform (GCP).

The container comes built with the following setting, which turns off support for GCP:

```
TF_NEED_GCP 0
```

### 6.1.3. TF_NEED_HDFS

The TF_NEED_HDFS parameter, as defined, disables support for the Hadoop
Distributed File System (HDFS).

The container comes built with the following setting, which turns off support for HDFS:

```
TF_NEED_HDFS 0
```

### 6.1.4. TF_ENABLE_XLA

The TF_ENABLE_XLA parameter, as defined, enables support for the Accelerated
Linear Algebra (XLA) backend.

The container comes built with the following setting, which turns on support for XLA:

```
TF_ENABLE_XLA 1
```
Chapter 7.
TENSORFLOW ENVIRONMENT VARIABLES

The following environment variable settings enable certain features within TensorFlow. They change and reduce the precision of the computation slightly and are enabled by default.

7.1. Added Or Modified Variables

In addition to the variables within the Dockerfile that are included in the Google TensorFlow container, the following variables have either been added or modified with the NVIDIA version of TensorFlow.

For variables not mentioned in this guide, see the Google documentation.

7.1.1. **TF_ADJUST_HUE_FUSED**

The *TF_ADJUST_HUE_FUSED* variable enables the use of fused kernels for the image hue.

This variable is enabled by default:

```bash
TF_ADJUST_HUE_FUSED 1
```

To disable the variable, run the following command:

```bash
export TF_ADJUST_HUE_FUSED=0
```

7.1.2. **TF_ADJUST_SATURATION_FUSED**

The *TF_ADJUST_SATURATION_FUSE* variable enables the use of fused kernels for the saturation adjustment.

This variable is enabled by default:

```bash
TF_ADJUST_SATURATION_FUSED 1
```

To disable the variable, run the following command:
7.1.3. ** TF_ENABLE_WINOGRAD_NONFUSED **

The `TF_ENABLE_WINOGRAD_NONFUSED` variable enables the use of the non-fused Winograd convolution algorithm.

This variable is enabled by default:

```
TF_ENABLE_WINOGRAD_NONFUSED 1
```

To disable the variable, run the following command:

```
export TF_ENABLE_WINOGRAD_NONFUSED=0
```

7.1.4. ** TF_AUTOTUNE_THRESHOLD **

The `TF_AUTOTUNE_THRESHOLD` variable improves the stability of the auto-tuning process used to select the fastest convolution algorithms. Setting it to a higher value improves stability, but requires a larger number of trial steps at the beginning of training before the best algorithms are found.

Within the container, this variable is set to the following:

```
export TF_AUTOTUNE_THRESHOLD=2
```

To set this variable to its default setting, run the following command:

```
export TF_AUTOTUNE_THRESHOLD=1
```

7.1.5. ** CUDA_DEVICE_MAX_CONNECTIONS **

The `CUDA_DEVICE_MAX_CONNECTIONS` variable solves performance issues related to streams on Tesla K80 GPUs.

Within the container, this variable is set to the following:

```
export CUDA_DEVICE_MAX_CONNECTIONS=12
```

To set this variable to its default setting, run the following command:

```
export CUDA_DEVICE_MAX_CONNECTIONS=8
```

7.1.6. ** TF_DISABLE_CUDNN_TENSOR_OP_MATH **

The `TF_DISABLE_CUDNN_TENSOR_OP_MATH` variable enables and disables Tensor Core math for cuDNN convolutions in TensorFlow. Tensor Core math is enabled by default, but can be disabled by setting this variable to 1. For more information, see Tensor Core Math.

This variable is disabled by default:

```
export TF_DISABLE_CUDNN_TENSOR_OP_MATH=0
```

To enable the variable, run the following command:
export TF_DISABLE_CUDNN_TENSOR_OP_MATH=1

### 7.1.7. **TF_DISABLE_CUDNN_RNN_TENSOR_OP_MATH**

The `TF_DISABLE_CUDNN_RNN_TENSOR_OP_MATH` variable enables and disables Tensor Core math for cuDNN RNNs in TensorFlow. Tensor Core math is enabled by default, but can be disabled by setting this variable to 1. For more information, see [Tensor Core Math](#).

This variable is disabled by default:

```
export TF_DISABLE_CUDNN_RNN_TENSOR_OP_MATH=0
```

To enable the variable, run the following command:

```
export TF_DISABLE_CUDNN_RNN_TENSOR_OP_MATH=1
```

### 7.1.8. **TF_DISABLE_CUBLAS_TENSOR_OP_MATH**

The `TF_DISABLE_CUBLAS_TENSOR_OP_MATH` variable enables and disables Tensor Core math for cuBLAS convolutions in TensorFlow. Tensor Core math is enabled by default, but can be disabled by setting this variable to 1. For more information, see [Tensor Core Math](#).

This variable is disabled by default:

```
export TF_DISABLE_CUBLAS_TENSOR_OP_MATH=0
```

To enable the variable, run the following command:

```
export TF_DISABLE_CUBLAS_TENSOR_OP_MATH=1
```

### 7.1.9. **TF_ENABLE_TENSOR_OP_MATH_FP32**

The `TF_ENABLE_TENSOR_OP_MATH_FP32` variable enables and disables Tensor Core math for float32 Matrix Multiplication operations in TensorFlow. Tensor Core math for float32 operations is disabled by default, but can be enabled by setting this variable to 1. For more information, see [Tensor Core Math](#).

This variable is disabled by default:

```
export TF_ENABLE_TENSOR_OP_MATH_FP32=0
```

To enable this variable, run the following command:

```
export TF_ENABLE_TENSOR_OP_MATH_FP32=1
```

### 7.1.10. **TF_DISABLE_NVTX_RANGES**

The `TF_DISABLE_NVTX_RANGES` variable enables and disables NVTX ranges in TensorFlow. NVTX ranges add operation name annotations to the execution timeline when profiling an application with Nsight Systems or the NVIDIA Visual Profiler. These NVTX ranges are enabled by default, but can be disabled by setting this variable to...
1. For more information on NVTX, see CUDA Toolkit Documentation: NVIDIA Tools Extension.

This variable is disabled by default:

```bash
export TF_DISABLE_NVTX_RANGES=0
```

To enable this variable, run the following command:

```bash
export TF_DISABLE_NVTX_RANGES=1
```
Chapter 8.
PERFORMANCE

To achieve optimum TensorFlow performance, for image based training, the container includes sample scripts that demonstrate efficient training of convolutional neural networks. The sample scripts may need to be modified to fit your application. The scripts can be found in the `/opt/tensorflow/nvidia-examples/cnn/` directory. Along with the training scripts, there is also some documentation that can be found in the `/opt/tensorflow/nvidia-examples/cnn/README.md` directory.

8.1. Tensor Core Math

The TensorFlow container includes support for the Volta architecture’s Tensor Cores, available on Tesla V100 GPUs. Tensor Cores deliver up to 12x higher peak TFLOPs for training. The container enables Tensor Core math by default; therefore, any models containing convolutions or matrix multiplies using the `tf.float16` data type will automatically take advantage of Tensor Core hardware whenever possible.

Tensor Core math can also be enabled for `tf.float32` matrix multiply operations by setting the `TF_ENABLE_TENSOR_OP_MATH_FP32` environment variable. This mode causes data to be internally reduced to float16 precision, which may affect training convergence.

For more information about the Volta architecture, see: Inside Volta: The World’s Most Advanced Data Center GPU.

8.1.1. Float16 Training

Training with reduced precision can in some cases lead to poor or unstable convergence. NVIDIA recommends the following strategies to minimize the effects of reduced precision during training (see `nvidia-examples/cnn/nvcnn.py` for a complete demonstration of float16 training):

1. Keep trainable variables in float32 precision and cast them to float16 before using them in the model. For example:

   ```python
   tf.cast(tf.get_variable(..., dtype=tf.float32), tf.float16)
   ```
2. Apply loss-scaling if the model struggles or fails to converge. Loss scaling involves multiplying the loss by a scale factor before computing gradients and then dividing the resulting gradients by the same scale again to re-normalize them. A typical loss scale factor for recurrent neural network models is 128. For example:

```python
loss, params = ...
scale = 128
grads = [grad / scale for grad in tf.gradients(loss * scale, params)]
```
Chapter 9.
TROUBLESHOOTING

9.1. Support

For more information about TensorFlow, including tutorials, documentation, and examples, see:

- TensorFlow tutorials
- TensorFlow API

For the latest TensorFlow Release Notes, see the Deep Learning Documentation website.
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 cuBLAS, CUDA, cuDNN, cuFFT, cuSPARSE, DIGITS, DGX, DGX-1, DGX Station, GRID, Jetson, Kepler, NVIDIA GPU Cloud, Maxwell, NCCL, NVLink, Pascal, Tegra, TensorRT, Tesla and Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the Unites States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2018 NVIDIA Corporation. All rights reserved.

www.nvidia.com