cuDNN Best Practices

BPG-09678-029_v8.1.0 | January 2021

Table of Contents

Chapter 1. Introduction.. 1
Chapter 2. Best Practices For Medical Imaging.. 2
2.1. Recommended Settings In cuDNN While Performing 3D Convolutions.. 2
2.1.1. cuDNN 8.x.x.. 2
2.1.2. cuDNN 7.6.x.. 3
Chapter 3. Medical Imaging Performance.. 6
Chapter 4. Medical Imaging Limitations.. 7

List of Tables

Table 1. Average speed-up of unique cuDNN (version 8.1.0 compared to 7.6.5) 3D convolution API calls on V100 and A100 for both FP16 and FP32........ 6

Chapter 1. Introduction

! **ATTENTION:** These guidelines are applicable to 3D convolution and deconvolution functions starting in NVIDIA® CUDA® Deep Neural Network library™ (cuDNN) v7.6.3.

This document provides guidelines for setting the cuDNN library parameters to enhance the performance of 3D convolutions. Specifically, these guidelines are focused on settings such as filter sizes, padding and dilation settings. Additionally, an application-specific use-case, namely, medical imaging, is presented to demonstrate the performance enhancement of 3D convolutions with these recommended settings.

Specifically, these guidelines are applicable to the following functions and their associated data types:

- cudnnConvolutionForward()
- cudnnConvolutionBackwardData()
- cudnnConvolutionBackwardFilter()

For more information, see the cuDNN Developer Guide and cuDNN API.

Chapter 2. Best Practices For Medical Imaging

To optimize your performance in your model, ensure you meet the following general guidelines:

Layout

The layout is in NCHW format.

Filter size

The filter size is `Tx1x1`, `Tx2x2`, `Tx3x3`, `Tx5x5`, where `T` is a positive integer. There are additional limits for the value of `T` in `wgrad` and strided `dgrad`.

Stride

Arbitrary for forward and backward filter; `dgrad`/`deconv`: 1x1x1 or 2x2x2 with 2x2x2 filter.

Dilation

The dilation is 1x1x1.

Platform

The platform is Volta, Turing, and Ampere with input/output channels divisible by 8.

Batch/image size

cuDNN will fallback to non-Tensor Core kernel if it determines that the workspace required is larger than 256MB of GPU memory. The workspace required depends on many factors. For the Tensor Core kernels, the workspace size generally scales linearly with output tensor size. Therefore, this can be mitigated by using smaller image sizes or minibatch sizes.

2.1. Recommended Settings In cuDNN While Performing 3D Convolutions

The following tables show the specific improvements that were made in each release.

2.1.1. cuDNN 8.x.x

Recommended settings while performing 3D convolutions for cuDNN 8.x.x.

		8.0.3 - 8.1.0	8.0.0 and 8.0.1 Preview - 8.0.2
Platform		NVIDIA Ampere GPU architecture NVIDIA Turing GPU architecture NVIDIA Volta GPU architecture	
Convolution (3D or 2D)		3D and 2D	

Convolution or deconvolution (`fprop`, `dgrad`, or `wgrad`)		`fprop`
		`dgrad`
		`wgrad`	

Grouped convolution	Yes or No	Yes	
	Group size	`C_per_group == K_per_group == {4,8,16,32,64,128,256}`	`C_per_group == K_per_group == {4,8,16,32}`
Data layout format (`NHWC`/`NCHW`)[1]		NDHWC	
Input/output precision (FP16, FP32, or FP64)		FP16 and FP32[2]	
Accumulator (compute) precision (FP16, FP32, or FP64)		FP32	
Filter (kernel) sizes		No limitation	
Padding		No limitation	
Image sizes		2GB limitation for a tensor	
Number of channels	C	`0 mod 8`	
	K	`0 mod 8`	
Convolution mode		Cross-correlation and convolution	
Strides		`dgrad`: 1x1x1 or 2x2x2	
Dilation		No limitation	
Data pointer alignment		All data pointers are 16-bytes aligned.	

2.1.2. cuDNN 7.6.x

Recommended settings while performing 3D convolutions for cuDNN 7.6.x.

	7.6.5	7.6.4	7.6.2	7.6.1
Platform	Turing Volta	Volta		
Convolution (3D or 2D)	3D and 2D	3D		
Convolution or deconvolution (`fprop`, `dgrad`, or `wgrad`)	`fprop` `dgrad`		`fprop` `dgrad`	`fprop` `dgrad`

[1] `NHWC`/`NCHW` corresponds to `NDHWC`/`NCDHW` in 3D convolution.

[2] With `CUDNN_TENSOROP_MATH_ALLOW_CONVERSION` pre-Ampere. Default TF32 math in Ampere.

		7.6.5	7.6.4	7.6.2	7.6.1
		`wgrad`			`wgrad`
Grouped convolution	Yes or No	Yes		No	
	Group size	`C_per_group == K_per_group == {4,8,16,32}`		NA	
Data layout format (`NHWC`/`NCHW`)[3]		NCDHW			NCDHW[4]
Input/output precision (FP16, FP32, or FP64)		FP16		FP16 or FP32	FP16[5] or FP32[6]
Accumulator (compute) precision (FP16, FP32, or FP64)		FP32		Better to be the same with input/output precision.	FP32
Filter (kernel) sizes		2x2x2 T[7]x1x1 Tx2x2 Tx3x3 Tx5x5			1x1x1 2x2x2 3x3x3 5x5x5 Tx1x1 Tx2x2 Tx3x3 Tx5x5 Tx1x1 Tx2x2 Tx3x3 Tx5x5
Padding		No limitation			Filter // 2[8]
Image sizes		256 MB WS limit		No limitation	256 MB WS limit
Number of channels	C	Arbitrary			0 mod 8
	K	Arbitrary			0 mod 8
Convolution mode		Cross-correlation for `dgrad`; otherwise, both modes		No limitation Cross-correlation	
Strides		1x1x1 and 2x2x2 strides for `dgrad`		2x2x2	1x1x1

[3] `NHWC`/`NCHW` corresponds to `NDHWC`/`NCDHW` in 3D convolution.

[4] With NCHW <> NHWC format transformation.

[5] FP16: `CUDNN_TENSOROP_MATH`

[6] FP32: `CUDNN_TENSOROP_MATH_ALLOW_CONVERSION`

[7] An arbitrary positive value.

[8] `padding = filter // 2`

	7.6.5	7.6.4	7.6.2	7.6.1
			Arbitrary stride	
Dilation	1x1x1			

Chapter 3. Medical Imaging Performance

The following table shows the average speed-up of **unique cuDNN 3D convolution calls** for each network on V100 and A100 GPUs that satisfies the conditions in Best Practices For Medical Imaging. The end-to-end training performance will depend on a number of factors, such as framework overhead, kernel run time, and model architecture type.

Table 1. Average speed-up of unique cuDNN (version 8.1.0 compared to 7.6.5) 3D convolution API calls on V100 and A100 for both FP16 and FP32.

		V100 **(8.1.0 vs. 7.6.5)**		**A100** **(8.1.0 vs. 7.6.5)**	
Model	**Batchsize**	**FP16**	**FP32**	**FP16**	**FP32**
V-Net (3D-Image segmentation)	2	2.2x	2.4x	2.4x	7.3x
	8	2.3x	1.8x	3.4x	5.3x
	16	2.3x	2.1x	3.9x	6x
	32	3.1x	1.6x	5.5x	4.4x
3D-UNet (3D-Image Segmentation)	2	3.5x	1.5x	7.3x	6.4x
	4	5x	1.6x	11.2x	2.6x

Chapter 4. Medical Imaging Limitations

Your application will be functional but slow if the model has:

- Channel counts lower than 32 (gets worse the lower it is)
- Data gradients for convolutions with stride

If the above is in the network, use `cuDNNFind` to get the best option.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform 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 document. NVIDIA accepts no 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 document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

ARM

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.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
www.nvidia.com

Trademarks

NVIDIA, the NVIDIA logo, and cuBLAS, CUDA, CUDA Toolkit, cuDNN, DALI, DIGITS, DGX, DGX-1, DGX-2, DGX Station, DLProf, GPU, JetPack, Jetson, Kepler, Maxwell, NCCL, Nsight Compute, Nsight Systems, NVCaffe, NVIDIA Ampere GPU architecture, NVIDIA Deep Learning SDK, NVIDIA Developer Program, NVIDIA GPU Cloud, NVLink, NVSHMEM, PerfWorks, Pascal, SDK Manager, T4, 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.

Copyright

© 2019-2021 NVIDIA Corporation. All rights reserved.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
www.nvidia.com