TABLE OF CONTENTS

Chapter 1. Introduction........................................................................................................... 1
Chapter 2. Mixed Precision Training....................................................................................... 2
  2.1. Half Precision Format........................................................................................................ 3
  2.2. Tensor Core Math............................................................................................................. 4
  2.3. Considering When Training With Mixed Precision......................................................... 4
    2.3.1. Loss Scaling To Preserve Small Gradient Magnitudes................................................ 6
    2.3.2. Choosing A Scaling Factor....................................................................................... 7
Chapter 3. Automatic Mixed Precision................................................................................ 9
Chapter 4. Optimizing For Tensor Cores................................................................................ 10
  4.1. Satisfying Tensor Core Shape Constraints........................................................................ 10
  4.2. Increasing Arithmetic Intensity..................................................................................... 11
  4.3. Decreasing Non-Tensor Core Work.................................................................................. 11
Chapter 5. Multi-GPU Training............................................................................................ 12
Chapter 6. Prerequisites......................................................................................................... 13
Chapter 7. Frameworks......................................................................................................... 14
Chapter 8. Deploying DNNs.................................................................................................. 26
Chapter 9. FAQs.................................................................................................................. 27
Chapter 1.
INTRODUCTION

There are numerous benefits to using numerical formats with lower precision than 32-bit floating point. First, they require less memory, enabling the training and deployment of larger neural networks. Second, they require less memory bandwidth, thereby speeding up data transfer operations. Third, math operations run much faster in reduced precision, especially on GPUs with tensor core support for that precision. Mixed precision training achieves all these benefits while ensuring that no task-specific accuracy is lost compared to full precision training. It does so by identifying the steps that require full precision and using 32-bit floating point for only those steps while using 16-bit floating point everywhere else.
Chapter 2.
MIXED PRECISION TRAINING

**Mixed precision** training offers significant computational speedup by performing operations in half-precision format, while storing minimal information in single-precision to retain as much information as possible in critical parts of the network. Since the introduction of tensor cores in the Volta and Turing architectures, significant training speedups are experienced by switching to mixed precision -- up to 3x overall speedup on the most arithmetically intense model architectures. Using mixed precision training requires two steps:

1. Porting the model to use the FP16 data type where appropriate.
2. Adding loss scaling to preserve small gradient values.

The ability to train deep learning networks with lower precision was introduced in the Pascal architecture and first supported in CUDA® 8 in the NVIDIA Deep Learning SDK. **Mixed precision** is the combined use of different numerical precisions in a computational method.

**Half precision** (also known as FP16) data compared to higher precision FP32 vs FP64 reduces memory usage of the neural network, allowing training and deployment of larger networks, and FP16 data transfers take less time than FP32 or FP64 transfers.

**Single precision** (also known as 32-bit) is a common floating point format (**float** in C-derived programming languages), and 64-bit, known as double precision (**double**).

Deep Neural Networks (DNNs) have led to breakthroughs in a number of areas, including image processing and understanding, language modeling, language translation, speech processing, game playing, and many others. DNN complexity has been increasing to achieve these results, which in turn has increased the computational resources required to train these networks. One way to lower the required resources is to use lower-precision arithmetic, which has the following benefits.

**Decrease the required amount of memory**

Half-precision floating point format (FP16) uses 16 bits, compared to 32 bits for single precision (FP32). Lowering the required memory enables training of larger models or training with larger mini-batches.
Shorten the training or inference time
Execution time can be sensitive to memory or arithmetic bandwidth. Half-precision halves the number of bytes accessed, thus reducing the time spent in memory-limited layers. NVIDIA GPUs offer up to 8x more half precision arithmetic throughput when compared to single-precision, thus speeding up math-limited layers.

Figure 1 Training curves for the bigLSTM English language model shows the benefits of the mixed-precision training techniques. The Y-axis is training loss. Mixed precision without loss scaling (grey) diverges after a while, whereas mixed precision with loss scaling (green) matches the single precision model (black).

Since DNN training has traditionally relied on IEEE single-precision format, this guide will focus on how to train with half precision while maintaining the network accuracy achieved with single precision (as Figure 1). This technique is called mixed-precision training since it uses both single- and half-precision representations.

2.1. Half Precision Format

IEEE 754 standard defines the following 16-bit half-precision floating point format: 1 sign bit, 5 exponent bits, and 10 fractional bits.

Exponent is encoded with 15 as the bias, resulting [-14, 15] exponent range (two exponent values, 0 and 31, are reserved for special values). An implicit lead bit 1 is assumed for normalized values, just like in other IEEE floating point formats.

Half precision format leads to the following dynamic range and precision:

**Normalized values**
- $2^{14}$ to $2^{15}$, 11 bits of significand

**Denormal values**
- $2^{-24}$ to $2^{-15}$, significand bits decrease as the exponent gets smaller. Exponent k in [-24, -15] range results in (25 - k) bits of significand precision.
Some example magnitudes:

**Maximum normalized**
65,504

**Minimum normalized**
$2^{-14} \approx -6.10e-5$

**Minimum denormal**
$2^{-24} \approx 5.96e-8$

Half precision dynamic range, including denormals, is 40 powers of 2. For comparison, single precision dynamic range including denormals is 264 powers of 2.

### 2.2. Tensor Core Math

The Volta generation of GPUs introduces tensor cores, which provide 8x more throughput than single precision math pipelines. Each tensor core performs $D = A \times B + C$, where $A$, $B$, $C$ and $D$ are matrices. $A$ and $B$ are half precision 4x4 matrices, whereas $D$ and $C$ can be either half or single precision 4x4 matrices. In other words, tensor core math can accumulate half precision products into either single or half precision outputs.

In practice, higher performance is achieved when $A$ and $B$ dimensions are multiples of 8. cuDNN v7 and cuBLAS 9 include some functions that invoke tensor core operations, for performance reasons these require that input and output feature map sizes are multiples of 8. For more information, see the [cuDNN Developer Guide](#).

The reason half precision is so attractive is that the V100 GPU has 640 tensor cores, so they can all be performing 4x4 multiplications all at the same time. The theoretical peak performance of the tensor cores on the V100 is approximately 120 TFLOPS. This is about an order of magnitude (10x) faster than double precision (FP64) and about 4 times faster than single precision (FP32).

Matrix multiplies are at the core of Convolutional Neural Networks (CNN). CNN’s are very common in deep learning in many networks. Beginning in CUDA 9 and cuDNN 7, the convolution operations are done using tensor cores whenever possible. This can greatly improve the training speed as well as the inference speed of CNN’s or models that contain convolutions.

### 2.3. Considering When Training With Mixed Precision

Given a framework that supports tensor core math, many networks can be trained faster by simply enabling the tensor core path in the framework (choosing FP16 format for tensors and/or convolution/fully-connected layers; for more details see Frameworks) and keeping all the hyperparameters of FP32 training session.

However, some networks require their gradient values to be shifted into FP16 representable range to match the accuracy of FP32 training sessions. The figure below illustrates one such case.
Figure 2  Histogram of activation gradient magnitudes throughout FP32 training of Multibox SSD network. The x-axis is logarithmic, except for the zero entry. For example, 66.8% of values were 0, 4% had magnitude in the \(2^{-32} \text{ to } 2^{-30}\) range.

However, this isn’t always the case. You may have to do some scaling and normalization to use FP16 during training.

Figure 3  Histogram of activation gradient magnitudes throughout FP32 training of Multibox SSD network. Both x- and y-axes are logarithmic.
Consider the histogram of activation gradient values (shown with linear and log y-scales above), collected across all layers during FP32 training of Multibox SSD detector network (VGG-D backbone). When converted to FP16, 31% of these values become zeros, leaving only 5.3% as non-zeros which for this network lead to divergence during training.

Much of the FP16 representable range was left unused by the gradient values. Therefore, if we shift the gradient values to occupy more of that range, we can preserve many values that are otherwise lost to 0s.

For this particular network, shifting by 3 exponent values (multiply by 8) was sufficient to match the accuracy achieved with FP32 training by recovering the relevant values lost to 0. Shifting by 15 exponent values (multiplying by 32K) would recover all but 0.1% of values lost to 0 when converting to FP16 and still avoid overflow. In other words, FP16 dynamic range is sufficient for training, but gradients may have to be scaled to move them into the range to keep them from becoming zeros in FP16.

### 2.3.1. Loss Scaling To Preserve Small Gradient Magnitudes

As was shown in the previous section, successfully training some networks requires gradient value scaling to keep them from becoming zeros in FP16. This can be efficiently achieved with a single multiplication by scaling the loss values computed in the forward pass, prior to starting backpropagation. By the chain rule, backpropagation ensures that all the gradient values are scaled by the same amount. This requires no extra operations during backpropagation and keeps the relevant gradient values from becoming zeros and losing that gradient information.

Weight gradients must be unscaled before weight update, to maintain the magnitude of updates the same as in FP32 training. It is simplest to perform this descaling right after the backward pass but before gradient clipping or any other gradient-related computations. This ensures that no hyperparameters (such as gradient clipping threshold, weight decay, etc.) have to be adjusted.

While many networks match FP32 training results when all tensors are stored in FP16, some require updating an FP32 copy of weights. Furthermore, values computed by large reductions should be left in FP32. Examples of this include statistics (mean and variance) computed by batch-normalization, SoftMax.

Batch-normalization can still take FP16 inputs and outputs, saving half the bandwidth compared to FP32, it’s just that the statistics and value adjustment should be done in FP32. This leads to the following high-level procedure for training:

1. Maintain a master copy of weights in FP32
2. For each iteration:
   a. Make an FP16 copy of the weights
   b. Forward propagation (FP16 weights and activations)
   c. Multiply the resulting loss with the scaling factor $S$
   d. Backward propagation (FP16 weights, activations, and their gradients)
2.3.2. Choosing A Scaling Factor

The procedure described in the previous section requires you to pick a loss scaling factor to adjust the gradient magnitudes. There is no downside to choosing a large scaling factor as long as it doesn’t cause overflow during backpropagation, which would lead to weight gradients containing infinities or NaNs, that in turn would irreversibly damage the weights during the update. These overflows can be easily and efficiently detected by inspecting the computed weight gradients, for example, multiply the weight gradient with 1/S step in the previous section. One option is to skip the weight update when an overflow is detected and simply move on to the next iteration.

There are several options to choose the loss scaling factor. The simplest one is to pick a constant scaling factor. We trained a number of feed-forward and recurrent networks with Tensor Core math for various tasks with scaling factors ranging from 8 to 32K (many networks did not require a scaling factor), matching the network accuracy achieved by training in FP32. However, since the minimum required scaling factor can depend on the network, framework, minibatch size, etc., some trial and error may be required when picking a scaling value. A constant scaling factor can be chosen more directly if gradient statistics are available. Choose a value so that its product with the maximum absolute gradient value is below 65,504 (the maximum value representable in FP16).

A more robust approach is to choose the loss scaling factor dynamically. The basic idea is to start with a large scaling factor and then reconsider it in each training iteration. If no overflow occurs for a chosen number of iterations N then increase the scaling factor. If an overflow occurs, skip the weight update and decrease the scaling factor. We found that as long as one skips updates infrequently the training schedule does not have to be adjusted to reach the same accuracy as FP32 training. Note that N effectively limits how frequently we may overflow and skip updates. The rate for scaling factor update can be adjusted by picking the increase/decrease multipliers as well as N, the number of non-overflow iterations before the increase. We successfully trained networks with N = 2000, increasing scaling factor by 2, decreasing scaling factor by 0.5, many other settings are valid as well. Dynamic loss-scaling approach leads to the following high-level training procedure:

1. Maintain a master copy of weights in FP32.
2. Initialize \( S \) to a large value.
3. For each iteration:
   a. Make an FP16 copy of the weights.
   b. Forward propagation (FP16 weights and activations).
   c. Multiply the resulting loss with the scaling factor \( S \).
   d. Backward propagation (FP16 weights, activations, and their gradients).
   e. If there is an Inf or NaN in weight gradients:
      a. Reduce \( S \).
      b. Skip the weight update and move to the next iteration.
f. Multiply the weight gradient with 1/S.
g. Complete the weight update (including gradient clipping, etc.).
h. If there hasn’t been an Inf or NaN in the last N iterations, increase S.
Using mixed precision training requires three steps:

1. Converting the model to use the float16 data type where possible.
2. Keeping float32 master weights to accumulate per-iteration weight updates.
3. Using loss scaling to preserve small gradient values.

Frameworks that support fully automated mixed precision training also support:

• Automatic loss scaling and master weights integrated into optimizer classes
• Automatic casting between float16 and float32 to maximize speed while ensuring no loss in task-specific accuracy

In those frameworks with automatic support, using mixed precision can be as simple as adding one line of code or enabling a single environment variable. Currently, the frameworks with support for automatic mixed precision are TensorFlow, PyTorch, and MXNet. See Automatic Mixed Precision for Deep Learning for more information, along with the Frameworks section below.
Chapter 4.
OPTIMIZING FOR TENSOR CORES

NVIDIA tensor cores provide hardware acceleration for mixed precision training. On a
V100 GPU, tensor cores can speed up matrix multiply and convolution operations by up
to 8x in float16 over their float32 equivalents.

Taking full advantage of tensor cores may require changes to model code. This section
describes three steps you can take to maximize the benefit that tensor cores provide:

1. Satisfy tensor core shape constraints
2. Increase arithmetic intensity
3. Decrease fraction of work in non-tensor core operations

The above benefits are ordered by increasing complexity, and in particular, the
first step (satisfying shape constraints) usually provides most of the benefit for little
effort.

4.1. Satisfying Tensor Core Shape Constraints

Due to their design, tensor cores have shape constraints on their inputs.

For matrix multiplication:
- On FP16 inputs, all three dimensions (M, N, K) must be multiples of 8.
- On INT8 inputs (Turing only), all three dimensions must be multiples of 16.

For convolution:
- On FP16 inputs, input and output channels must be multiples of 8.
- On INT8 inputs (Turing only), input and output channels must be multiples of 16.

In practice, for mixed precision training, our recommendations are:

1. Choose mini-batch to be a multiple of 8
2. Choose linear layer dimensions to be a multiple of 8
3. Choose convolution layer channel counts to be a multiple of 8
4. For classification problems, pad vocabulary to be a multiple of 8
5. For sequence problems, pad the sequence length to be a multiple of 8

4.2. Increasing Arithmetic Intensity

Arithmetic intensity is a measure of how much computational work is to be performed in a kernel per input byte. For example, a V100 GPU has 125 TFLOPs of math throughput and 900 GB/s of memory bandwidth. Taking the ratio of the two, we see that any kernel with fewer than ~140 FLOPs per input byte will be memory-bound. That is, tensor cores cannot run at full throughput because memory bandwidth will be the limiting factor. A kernel with sufficient arithmetic intensity to allow full tensor core throughput is compute-bound.

It is possible to increase arithmetic intensity both in model implementation and model architecture.

To increase arithmetic intensity in model implementation:

- Concatenate weights and gate activations in recurrent cells.
- Concatenate activations across time in sequence models.

To increase arithmetic intensity in model architecture:

- Prefer dense math operations.
  - For example, vanilla convolutions have much higher arithmetic intensity than depth-wise separable convolutions.
- Prefer wider layers when possible accuracy-wise.

4.3. Decreasing Non-Tensor Core Work

Many operations in deep neural networks are not accelerated by tensor cores, and it is important to understand the effect this has on end-to-end speedups. For example, suppose a model spends one half of the total training time in tensor core-accelerated operations (matrix multiplication and convolution). If tensor cores provide a 5x speedup for those operations, then the total speedup will be $1 / \left( \frac{1}{2} + \frac{1}{10} \right) = 1.67x$.

In general, as tensor core operations represent a decreasing fraction of total work, the more important it is to focus on optimizing non-tensor core operations. It is possible to speedup up these operations by hand, using custom CUDA implementations along with framework integration. Furthermore, frameworks are beginning to provide support for automatically speeding up non-tensor core ops with compiler tools. Examples include XLA for TensorFlow and the PyTorch JIT.
Chapter 5.
MULTI-GPU TRAINING

For multi-GPU training, the same strategy applies for loss scaling. NCCL supports both half precision floats and normal floats, therefore, a developer can choose which precision they want to use to aggregate gradients. Batch size considerations depend on your training framework.
Chapter 6.
PREREQUISITES

To take advantage of mixed precision training, ensure you meet the following minimum requirements:

1. Run on the Volta or Turing architecture.
2. Install an NVIDIA Driver. Currently CUDA 10.1 is supported, which requires NVIDIA driver release 418.xx+. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you may use NVIDIA driver release 384.111+ or 410. The CUDA driver’s compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.
3. Install the CUDA® Toolkit™.
4. Install cuDNN.

If using an NVIDIA optimized framework container, that was pulled from the NGC container registry, you will still need to install an NVIDIA driver on your base operating system. However, CUDA and cuDNN will come included in the container. For more information, see the Frameworks Support Matrix.
Most major deep learning frameworks have begun to merge support for half precision training techniques that exploit tensor core calculations in Volta and Turing. Additional optimization pull requests are at various stages and listed in their respective section.

For NVCaffe, Caffe2, MXNet, Microsoft Cognitive Toolkit, PyTorch, TensorFlow and Theano, tensor core acceleration is automatically enabled if FP16 storage is enabled. While frameworks like Torch will tolerate the latest architecture, it currently does not exploit tensor core functionality.

### 7.1. PyTorch

PyTorch includes support for FP16 storage and tensor core math. To achieve optimum performance, you can train a model using tensor core math and mixed precision.

#### 7.1.1. Automatic Mixed Precision Training In PyTorch

The automatic mixed precision feature is available starting inside the NVIDIA NGC PyTorch 19.03 container.

To get started, we recommend using AMP (Automatic Mixed Precision), which enables mixed precision in only 3 lines of Python. AMP is available through NVIDIA's Apex repository of mixed precision and distributed training tools. The AMP API is documented in detail here.

#### 7.1.2. Success Stories

The models where we have seen speedup using mixed precision are:
### Table 1  Mixed precision model speedup

<table>
<thead>
<tr>
<th>Model</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA Sentiment Analysis</td>
<td>4.5X speedup</td>
</tr>
<tr>
<td>FAIRSeq</td>
<td>3.5X speedup</td>
</tr>
<tr>
<td>GNMT</td>
<td>2X speedup</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>2X speedup</td>
</tr>
</tbody>
</table>

### 7.1.3. Tensor Core Optimized Model Scripts For PyTorch

The tensor core examples provided in GitHub focus on achieving the best performance and convergence from NVIDIA Volta tensor cores by using the latest deep learning example networks and model scripts for training.

Each example model trains with mixed precision tensor cores on Volta, therefore you can get results much faster than training without tensor cores. This model is tested against each NGC monthly container release to ensure consistent accuracy and performance over time. The PyTorch container includes the following PyTorch tensor core examples:

- An implementation of the Mask R-CNN model. Mask R-CNN is a convolution based neural network for the task of object instance segmentation. The paper describing the model can be found [here](#). NVIDIA’s Mask R-CNN model is an optimized version of Facebook's implementation, leveraging mixed precision arithmetic using tensor cores on NVIDIA Tesla V100 GPUs for 1.3x faster training time while maintaining target accuracy.

- An implementation of the Tacotron 2 and WaveGlow v1.1 model. This text-to-speech (TTS) system is a combination of two neural network models: a modified Tacotron 2 model from the Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions paper and a flow-based neural network model from the WaveGlow: A Flow-based Generative Network for Speech Synthesis paper.

- An implementation of the SSD300 v1.1 model. The SSD300 v1.1 model is based on the SSD: Single Shot MultiBox Detector paper. The main difference between this model and the one described in the paper is in the backbone. Specifically, the VGG model is obsolete and is replaced by the ResNet50 model.

- An implementation of the Neural Collaborative Filtering (NCF) model. The NCF model focuses on providing recommendations, also known as collaborative filtering; with implicit feedback. The training data for this model should contain binary information about whether a user interacted with a specific item. NCF was first described by Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu and Tat-Seng Chua in the Neural Collaborative Filtering paper.

- An implementation of the Transformer model architecture. The Transformer model is based on the optimized implementation in Facebook's Fairseq NLP Toolkit and is built on top of PyTorch. The original version in the Fairseq project was developed using Tensor Cores, which provides significant training speedup. Our implementation improves the performance and is tested on a DGX-1 16GB.
- An implementation of the ResNet50 model. The ResNet50 v1.5 model is a modified version of the original ResNet50 v1 model.

### 7.1.4. Manual Conversion To Mixed Precision In PyTorch

We recommend using AMP to implement mixed precision in your model. However, if you wish to implement mixed precision yourself, refer to our GTC talk on manual mixed precision ([video](#), [slides](#)). An example of manual mixed precision for Imagenet training can be found [here](#), and an RNN example can be found [here](#).

### 7.2. TensorFlow

TensorFlow supports FP16 storage and tensor core math. Models that contain convolutions or matrix multiplications using the `tf.float16` data type will automatically take advantage of tensor core hardware whenever possible.

In order to make use of tensor cores, FP32 models will need to be converted to use a mix of FP32 and FP16. This can be done either automatically using automatic mixed precision (AMP) or manually.

#### 7.2.1. Automatic Mixed Precision Training In TensorFlow

The automatic mixed precision feature is available starting inside the [NVIDIA NGC TensorFlow 19.03 container](#). Enable this feature inside the container requires simply setting one environment variable:

```bash
export TF_ENABLE_AUTO_MIXED_PRECISION=1
```

You can also set the environment variable inside a TensorFlow Python script by calling the following at the beginning of the script:

```python
os.environ['TF_ENABLE_AUTO_MIXED_PRECISION'] = '1'
```

When enabled, automatic mixed precision will do two things:

1. Insert the appropriate cast operations into your TensorFlow graph to use float16 execution and storage where appropriate.
2. Turn on automatic loss scaling inside the training Optimizer object.

For more information on automatic mixed precision, see the TensorFlow User Guide.

#### 7.2.2. Success Stories

The models where we have seen speedup using mixed precision are:

<table>
<thead>
<tr>
<th>Model</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT Q&amp;A</td>
<td>3.3X speedup</td>
</tr>
</tbody>
</table>
7.2.3. Tensor Core Optimized Model Scripts For TensorFlow

The tensor core examples provided in GitHub focus on achieving the best performance and convergence from NVIDIA Volta tensor cores by using the latest deep learning example networks and model scripts for training.

Each example model trains with mixed precision tensor cores on Volta, therefore you can get results much faster than training without tensor cores. This model is tested against each NGC monthly container release to ensure consistent accuracy and performance over time. The TensorFlow container includes the following TensorFlow tensor core examples:

- An implementation of the SSD320 v1.2 model. The SSD320 v1.2 model is based on the SSD: Single Shot MultiBox Detector paper, which describes SSD as “a method for detecting objects in images using a single deep neural network”. Our implementation is based on the existing model from the TensorFlow models repository.

- An implementation of the Neural Collaborative Filtering (NCF) model. The NCF model is a neural network that provides collaborative filtering based on implicit feedback, specifically, it provides product recommendations based on user and item interactions. The training data for this model should contain a sequence of user ID, item ID pairs indicating that the specified user has interacted with, for example, was given a rating to or clicked on, the specified item.

- An implementation of the Bert model. BERT, or Bidirectional Encoder Representations from Transformers, is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks. This model is based on BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding paper. NVIDIA’s BERT is an optimized version of Google’s official implementation, leveraging mixed precision arithmetic and tensor cores on V100 GPUs for faster training times while maintaining target accuracy.

- An implementation of the U-Net Industrial Defect Segmentation model. This U-Net model is adapted from the original version of the U-Net model which is a convolutional auto-encoder for 2D image segmentation. U-Net was first introduced by Olaf Ronneberger, Philip Fischer, and Thomas Brox in the paper: U-Net: Convolutional Networks for Biomedical Image Segmentation. This work proposes a modified version of U-Net, called TinyUNet which performs efficiently and with very high accuracy on the industrial anomaly dataset DAGM2007.
An implementation of the GNMT v2 model. The GNMT v2 model is similar to the one discussed in the Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation paper. The most important difference between the two models is in the attention mechanism. In our model, the output from the first LSTM layer of the decoder goes into the attention module, then the re-weighted context is concatenated with inputs to all subsequent LSTM layers in the decoder at the current timestep.

An implementation of the ResNet-50 v1.5 model. The ResNet-50 v1.5 model is a modified version of the original ResNet-50 v1 model. The difference between v1 and v1.5 is in the bottleneck blocks which requires downsampling, for example, v1 has stride = 2 in the first 1x1 convolution, whereas v1.5 has stride = 2 in the 3x3 convolution. The following features were implemented in this model; data-parallel multi-GPU training with Horovod, tensor cores (mixed precision) training, and static loss scaling for tensor cores (mixed precision) training.

7.2.4. Manual Conversion To Mixed Precision Training In TensorFlow

1. Pull the latest TensorFlow container from the NVIDIA GPU Cloud (NGC) container registry. The container is already built, tested, tuned, and ready to run. The TensorFlow container includes the latest CUDA version, FP16 support, and is optimized for the latest architecture. For step-by-step pull instructions, see the Containers for Deep Learning Frameworks User Guide.

2. Use the `tf.float16` data type on models that contain convolutions or matrix multiplications. This data type automatically takes advantage of the tensor core hardware whenever possible, in other words, to increase your chances for tensor core acceleration, choose where possible multiple of eight linear layer matrix dimensions and convolution channel counts. For example:

   ```python
dtype = tf.float16
data = tf.placeholder(dtype, shape=(nbatch, nin))
weights = tf.get_variable('weights', (nin, nout), dtype)
biases  = tf.get_variable('biases',        nout,  dtype,
                         initializer=tf.zeros_initializer())
logits = tf.matmul(data, weights) + biases
```

3. Ensure the trainable variables are 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)
```

   This can also be achieved by using the `float32_variable_storage_getter` shown in the following example.

4. Ensure the SoftMax calculation is in float32 precision. For example:

   ```python
tf.losses.softmax_cross_entropy(target, tf.cast(logits, tf.float32))
```

5. Apply loss-scaling as outlined in the previous sections. 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. For example, to apply a constant loss scaling factor of 128:
7.3. MXNet

MXNet includes support for FP16 storage and tensor core math. To achieve optimum performance, you need to train a model using tensor core math and FP16 mode on MXNet.

The following procedure is typical for when you want to have your entire network in FP16. Alternatively, you can take output from any layer and cast it to FP16. Subsequent layers will be in FP16 and will use tensor core math if applicable.

7.3.1. Automatic Mixed Precision Training In MXNet

The automatic mixed precision feature is available starting inside the NVIDIA NGC MXNet 19.04 container.

Training deep learning networks is a very computationally intensive task. Novel model architectures tend to have increasing number of layers and parameters, which slows down training. Fortunately, new generations of training hardware as well as software optimizations, make it a feasible task.

However, where most of the (both hardware and software) optimization opportunities exists is in exploiting lower precision (like FP16) to, for example, utilize tensor cores available on new Volta and Turing GPUs. While training in FP16 showed great success in image classification tasks, other more complicated neural networks typically stayed in FP32 due to difficulties in applying the FP16 training guidelines.

That is where AMP (Automatic Mixed Precision) comes into play. It automatically applies the guidelines of FP16 training, using FP16 precision where it provides the most benefit, while conservatively keeping in full FP32 precision operations unsafe to do in FP16.

The MXNet AMP tutorial, located in the /opt/mxnet/nvidia-examples/AMP/AMP_tutorial.md directory of this container, shows you how to get started with mixed precision training using AMP for MXNet. As an example of a network we will use SSD network from GluonCV.

7.3.2. Tensor Core Optimized Model Scripts For MXNet

The tensor core examples provided in GitHub focus on achieving the best performance and convergence from NVIDIA Volta tensor cores by using the latest deep learning example networks and model scripts for training.

Each example model trains with mixed precision tensor cores on Volta, therefore you can get results much faster than training without tensor cores. This model is tested against each NGC monthly container release to ensure consistent accuracy and performance over time. The MXNet container includes the following MXNet tensor core examples:
An implementation of the ResNet50 model. The ResNet50 v1.5 model is a slightly modified version of the original ResNet50 v1 model that trains to a greater accuracy.

### 7.3.3. Manual Conversion To Mixed Precision Training In MXNet

1. Pull the latest MXNet container from the NVIDIA GPU Cloud (NGC) container registry. The container is already built, tested, tuned, and ready to run. The MXNet container includes the latest CUDA version, FP16 support, and is optimized for the latest architecture. For step-by-step pull instructions, see the Containers for Deep Learning Frameworks User Guide.

2. To use the IO pipeline, use the IndexedRecordIO format of input. It differs from the legacy RecordIO format, by including an additional index file with an .idx extension. The .idx file is automatically generated when using the im2rec.py tool, to generate new RecordIO files. If you already have the .rec file without the corresponding .idx file, you can generate the index file with tools/rec2idx.py tool:

   ```
   python tools/rec2idx.py <path to .rec file> <path to newly created .idx file>
   ```

3. To use FP16 training with MXNet, cast the data (input to the network) to FP16.

   ```
   mxnet.sym.Cast(data=input_data, dtype=numpy.float16)
   ```

4. Cast back to FP32 before SoftMax layer.

5. If you encounter precision problems, it is beneficial to scale the loss up by 128, and scale the application of the gradients down by 128. This ensures higher gradients during the backward pass calculation, but will still correctly update the weights. For example, if our last layer is `mx.sym.SoftmaxOutput` (cross-entropy loss), and the initial learning rate is 0.1, add a grad_scale parameter:

   ```
   mxnet.sym.SoftmaxOutput(other_args, grad_scale=128.0)
   ```

   When initializing the optimizer, rescale the gradient down prior to the application:

   ```
   mxnet.optimizer.SGD(other_args, rescale_grad=1.0/128)
   ```

   **Tip** When training in FP16, it is best to use multi-precision optimizers that keep the weights in FP32 and perform the backward pass in FP16. For example, for SGD with momentum, you would issue the following:

   ```
   mxnet.optimizer.SGD(other_args, momentum=0.9, multi_precision=True)
   ```

   Alternatively, you can pass `multi_precision`: True to the optimizer_params option in the model.fit method.

### 7.4. Caffe2
Caffe2 includes support for FP16 storage and tensor core math. To achieve optimum performance, you can train a model using tensor core math and FP16 mode on Caffe2.

When training a model on Caffe2 using tensor core math and FP16, the following actions need to take place:

- Prepare your data. You can generate data in FP32 and then cast it down to FP16. The GPU transforms path of the ImageInput operation can do this casting in a fused manner.
- Forward pass. Since data is given to the network in FP16, all of the subsequent operations will run in FP16 mode, therefore:
  - Select which operators need to have both FP16 and FP32 parameters by setting the type of Initializer used. Typically, the Conv and FC operators need to have both parameters.
  - Cast the output of forward pass, before SoftMax, back to FP32.
  - To enable Tensor Core, pass enable_tensor_core=True to ModelHelper when representing a new model.
  - Update the master FP32 copy of the weights using the FP16 gradients you just computed. For example:
    - Cast up gradients to FP32.
    - Update the FP32 copy of parameters.
    - Cast down the FP32 copy of parameters to FP16 for the next iteration.
- Gradient scaling.
  - To scale, multiply the loss by the scaling factor.
  - To descale, divide LR and weight_decay by the scaling factor.

7.4.1. Running FP16 Training On Caffe2

1. Pull the latest Caffe2 container from the NVIDIA GPU Cloud (NGC) container registry. The container is already built, tested, tuned, and ready to run. The Caffe2 container includes the latest CUDA version, FP16 support, and is optimized for the Volta architecture. For step-by-step pull instructions, see the Containers for Deep Learning Frameworks User Guide.

2. Run the following Python script with the appropriate command line arguments. You can test using the ResNet-50 image classification training script included in Caffe2.

```bash
caffe2/python/examples/resnet50_trainer.py --train_data <path> --test_data <path> --num-gpus <Int> --batch-size <Int> --dtype float16 --enable-tensor-core --cudnn_workspace_limit_mb 1024 --image_size 224
```

For more information about the additional command-line arguments, issue the following command:

```bash
caffe2/python/examples/resnet50_trainer.py --help
```

To enhance performance, the following changes must be made:
The network definition in `caffe2/python/models/resnet.py` must be changed to reflect version 1 of the network by changing the residual block striding from the 3x3 convolution to the first 1x1 convolution operator.

Enable optimized communication operators and disable some communication ops by adding the `use_nccl=True` and `broadcast_computed_params=False` flags to the `data_parallel_model.Parallelize` call in `caffe2/python/examples/resnet50_trainer.py`.

Add `decode_threads=3` and `use_gpu_transform=True` to the `brew.image_input` call. This tweaks the amount of CPU threads used for data decode and augmentation (value is per-GPU) and enables the use of the GPU for some data augmentation work.

Increase the number of host threads used to schedule operators on the GPUs by adding `train_model.net.Proto().num_workers = 4 * len(gpus)` after the call to `data_parallel_model.Parallelize`.

### 7.4.2. Caffe2 FP16 Example

For more information, you can find examples at: Caffe2 Python Examples.

### 7.5. Microsoft Cognitive Toolkit

Microsoft Cognitive Toolkit includes support for FP16 storage and tensor core math. To achieve optimum performance, you need to train a model using tensor core math and FP16 mode on Microsoft Cognitive Toolkit.

#### 7.5.1. Running FP16 Training On Microsoft Cognitive Toolkit

Tensor core math is turned on by default in FP16. The following procedure is typical of Microsoft Cognitive Toolkit using FP16 in a multi-layer perceptron MNIST example.

```python
import cntk as C
import numpy as np

input_dim = 784
num_output_classes = 10
num_hidden_layers = 1
hidden_layers_dim = 200

# Input variables denoting the features and label data
feature = C.input_variable(input_dim, np.float32)
label = C.input_variable(num_output_classes, np.float32)

feature16 = C.cast(feature, np.float16)
label16 = C.cast(label, np.float16)

with C.default_options(dtype=np.float16):
    # Instantiate the feedforward classification model
    scaled_input16 = C.element_times(C.constant(0.00390625, dtype=np.float16), feature16)
```

```python
```
z16 = C.layers.Sequential([C.layers.For(range(num_hidden_layers),
    lambda i: C.layers.Dense(hidden_layers_dim,
        activation=C.relu)), C.layers.Dense(num_output_classes)])
(scaled_input16)

c16 = C.cross_entropy_with_softmax(z16, label16)
pe16 = C.classification_error(z16, label16)

z = C.cast(z16, np.float32)
ce = C.cast(c16, np.float32)
pe = C.cast(pe16, np.float32)

# fake data with batch_size = 5
batch_size = 5
feature_data = np.random.randint(0, 256, (batch_size, 784)).astype(np.float32)
label_data = np.eye(num_output_classes)[np.random.randint(0, num_output_classes,
    batch_size)]

ce.eval({feature:feature_data, label:label_data})

7.5.2. Microsoft Cognitive Toolkit FP16 Example

For a more complete example of ResNet-50 with distributed training, see the
TrainResNet_ImageNet_Distributed.py example.

7.6. NVCaffe

NVCaffe includes support for FP16 storage and tensor core math. To achieve optimum
performance, you can train a model using tensor core math and FP16 mode on NVCaffe.

7.6.1. Running FP16 Training On NVCaffe

1. Pull the latest NVCaffe container from the NVIDIA GPU Cloud (NGC) container
registry. The container is already built, tested, tuned, and ready to run. The NVCaffe
container includes the latest CUDA version, FP16 support, and is optimized for the
latest architecture. For step-by-step pull instructions, if you have a DGX-1, see the
Containers for Deep Learning Frameworks User Guide, otherwise refer to the Using
NGC with Your NVIDIA TITAN or Quadro PC Setup Guide.

2. Experiment with the following training parameters:
   a) Before running the training script below, adjust the batch size for better
      performance. To do so, open the training settings with your choice of editor, for
      example, vim:

      ```
      caffe$ vim models/resnet50/train_val_fp16.prototxt
      And change the batch_size: 32 setting value to [64...128] * <Number of
      GPUs installed>.
      ```
   b) Experiment with pure FP16 mode by setting:

      ```
      default_forward_type: FLOAT16
      default_backward_type: FLOAT16
      default_forward_math: FLOAT16
      default_backward_math: FLOAT16
      ```
And by adding `solver_data_type: FLOAT16` to the file `models/resnet50/solver_fp16.prototxt`.

3. **Train ResNet-50.** Open:

```bash
caffe$ ./models/resnet50/train_resnet50_fp16.sh
```

When the training is finished, it should look similar to the following:

```
```

The performance number of 5268 img/sec was trained on an 8-GPU system. For a single GPU system, you could expect around 750 img/sec training with NVCaffe.

4. **View the output.** Issue the following command:

```bash
caffe$ python plot_top5.py -s models/resnet50/logs/resnet50_fp16.log
```

Your output should look similar to the following:

![ResNet-50 FP16 training log](image)

**Figure 4** ResNet-50 FP16 training log
7.6.2. NVCaffe FP16 Example

For examples on optimization, see the `models/resnet50/train_val_fp16.prototxt` file.

7.7. Theano

Theano includes support for FP16 storage and tensor core math. To make use of tensor core math, set the `dnn.conv.algo_XXX` configuration parameter to `time_once` or `time_on_shape_change`, for example:

```plaintext
[dnn]
conv.algo_fwd=time_once
conv.algo_bwd_filter=time_once
conv.algo_bwd_data=time_once
```

7.7.1. Running FP16 Training On Theano

Theano is fully parameterized on `floatX` type, therefore, to run most Theano scripts in FP16, you can issue:

```plaintext
THEANO_FLAGS="floatX=float16"
```
Chapter 8.
DEPLOYING DNNS

After you have trained a neural network, you can optimize and deploy the model for GPU inferencing with TensorRT™. For more information about optimizing and deploying using TensorRT, see Deep Learning SDK Documentation.
Q: Why am I not getting more performance?
A: Are you using FP16? If you’re not using NVIDIA containers, you need to rebuild your framework. In addition, there are a few extra steps you need to perform in order to use FP16. This document as well as the Deep Learning Performance Guide provides more details.

Q: What performance should I expect from V100 and mixed precision?
A: It is dependent on what you are doing. Usually, an SA will then attach some numbers with some data with ImageNet training performance.

Q: Is V100 optimized for Neural Machine?
A: This will differ based on your setup and environment. As an example, for OpenNMT performance, if your model is large enough, you should see about 1.5x performance improvement if all the work is on the GPU.

Q: What are other resources for how to use mixed precision?
A: Here are some additional resources that can help with understanding mixed precision:

- Mixed Precision Training (ICLR 2018).
- Mixed-Precision Training of Deep Neural Networks (NVIDIA Developer Blog).
- Mixed Precision Training of Deep Neural Networks (GTC 2019, S9143).
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, DALI, DIGITS, DGX, DGX-1, Jetson, Kepler, NVIDIA Maxwell, NCCL, NVLink, Pascal, Tegra, TensorRT, and Tesla 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

© 2019 NVIDIA Corporation. All rights reserved.

www.nvidia.com