MIXED PRECISION TRAINING

Michael O’Connor
MIXED PRECISION
What is the benefit?

Using mixed precision and Volta your networks can be:

1. 3-4x faster
2. Reduce memory consumption and bandwidth pressure
3. just as powerful

with no architecture change.
A MIXED PRECISION SOLUTION

Imprecise weight updates → "Master" weights in FP32

Gradients underflow → Loss (Gradient) Scaling

Maintain precision → Accumulate to FP32 (Tensor Cores)
MIXED SOLUTION: FP32 MASTER WEIGHTS

- **FP32 Master Weights**
- **FP32 Master Gradients**
- **FP16 Weights**
- **FP16 Loss**
- **FP16 Gradients**

Flowchart:
- Copy from FP32 Master Weights to FP16 Weights
- Forward Pass from FP16 Weights to FP16 Loss
- Backprop from FP16 Loss to FP16 Gradients
- Copy from FP32 Master Gradients to FP16 Gradients
- Apply from FP32 Master Weights to FP16 Weights
GRADIENTS RANGE OFFSET

![Diagram showing the range and offset of gradients in a histogram with logarithmic magnitude on the x-axis and percentage of all activation gradient values on the y-axis. The bars represent the distribution of gradient magnitudes, with a notable peak around 1/16. The diagram highlights the FP16 representable range, denormal (denoms) values, and the point at which gradients become zero in FP16.]
MIXED PRECISION TRAINING

FP32 Master Weights

Apply

Remove scale, (+clip, etc.)

Copy

Scaled FP32 Gradients

FP32 Gradients

FP16 Weights

Forward Pass

FP32 Loss

Loss Scaling

Scaled FP32 Loss

Scaled FP16 Gradients

Copy

Backprop
NVCAFFE V0.16 TRAINING ALEXNET

Single P100 GPU, Batch Size=128
RESNET-50 FP32 PERFORMANCE

Images per second

1 GPU  2 GPU  4 GPU  8 GPU

4/30/2017 : DGX-1 with Batch Size=64 per GPU. Chainer numbers are preliminary.
RESNET-50 MIXED PRECISION AND FP32

Images per second

- 1 GPU
- 2 GPU
- 4 GPU
- 8 GPU

MXNet FP32 GTC 2017
MXNet FP32 GTC 2018
MXNet Mixed GTC 2018
INFORMATION SOURCES
Where to learn about mixed precision training

CE8130 - Connect with the Experts: Deep Learning Training for Volta Tensor Cores Tu 2PM
S8923 - Training Neural Networks with Mixed Precision: Theory and Practice Wed 2PM
S81012 - Training Neural Networks with Mixed Precision: Real Examples Th 9 AM
CE8162 - Connect with the Experts: Deep Learning Training for Volta Tensor Cores Th 2PM

Mixed-Precision Training of Deep Neural Networks (NVIDIA Developer Blog)

Training with Mixed Precision (NVIDIA User Guide)