Using Texture Compression Hardware for Neural Network Inference

Hot Chips 2017

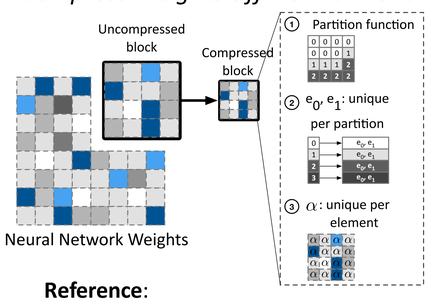
Using ASTC compression for DNN Weights

ASTC – Adaptive Scalable Texture Compression

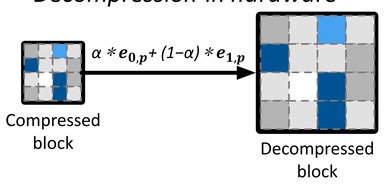
Qualities:
- Asymmetric – fast decode
- Hardware support in modern mobile GPUs
- Random access
- Flexible compression – from 8-bits to 0.22-bits per weight

Encoding:
- Fixed Encoded Block (128-bit)
- Variable Source Block Size – 4x4 up to 12x12 region
- Mechanism – endpoint definition with interpolation
- Partition Function – enables localized endpoints within a source block

Compress Weights Offline with ASTC



Decompression in hardware



Reference:

J. Nystad, A. Lassen, A. Pomianowski, S. Ellis, and T. Olson. Adaptive scalable texture compression. In Proceedings of the Fourth ACM SIGGRAPH / Eurographics Conference on High-Performance Graphics, EGGH-HPG'12, 2012.

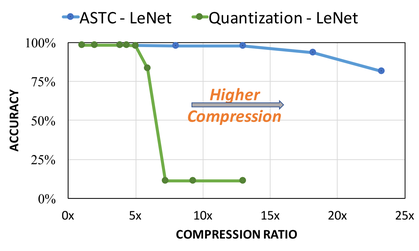
ASTC vs Weight Quantization (Fixed-Point)

Without re-training:

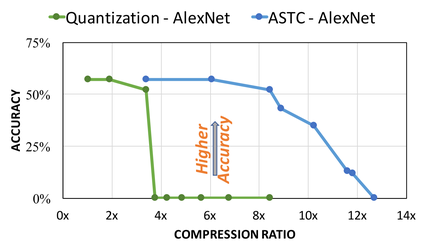
Compared to quantization, ASTC provides:
- Higher accuracy for fixed compression ratio
- Higher compression ratio for fixed accuracy
- Smooth trade-off between accuracy and compression ratio

Model (Dataset)	Original Size	No Loss in accuracy	
		Quant.	ASTC
LeNet (MNIST)	6.2 MB	**5.1x** (1.2 MB)	**13x** (0.47 MB)
AlexNet (ILSVRC12)	240 MB	**1.9x** (127 MB)	**6x** (39.5 MB)

Accuracy vs Compression Ratio LeNet (MNIST)



Accuracy vs Compression Ratio AlexNet (ILSVRC12)



Results on LeNet and AlexNet

With re-training:
- **18x** compression for LeNet and **8.7x** compression for AlexNet with no loss in accuracy
- Hardware support for ASTC available in most modern mobile GPUs
- Enable inference for state-of-the-art models in Edge devices

Model (Dataset)	Original Size	No Loss		<5% Loss	
		Quant.	ASTC	Quant.	ASTC
LeNet (MNIST)	6.2 MB	**5.1x** (1.2 MB)	**18.2** (0.34 MB)	**5.1x** (1.2 MB)	**23.2x** (0.27 MB)
AlexNet (ILSVRC12	240 MB	**1.9x** (127 MB)	**8.7x** (27.6 MB)	**3.4x** (70.8 MB)	**10.8x** (22.2 MB)

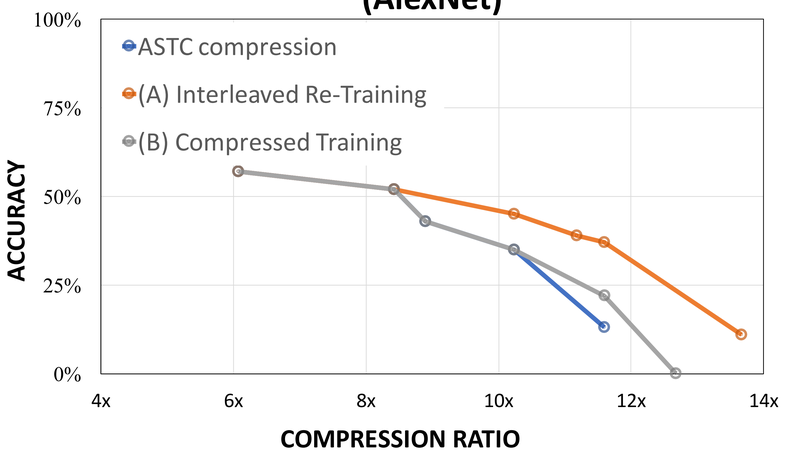
Re-Training ASTC Compressed Weights

Two techniques for re-training:

(A) Interleaving Re-Training: which interleaves re-training iterations with compression (Better accuracy)

(B) Compressed Training: which trains ($e_{0,p}$,$e_{1,p}$) and α for a fixed partition function (Faster training)

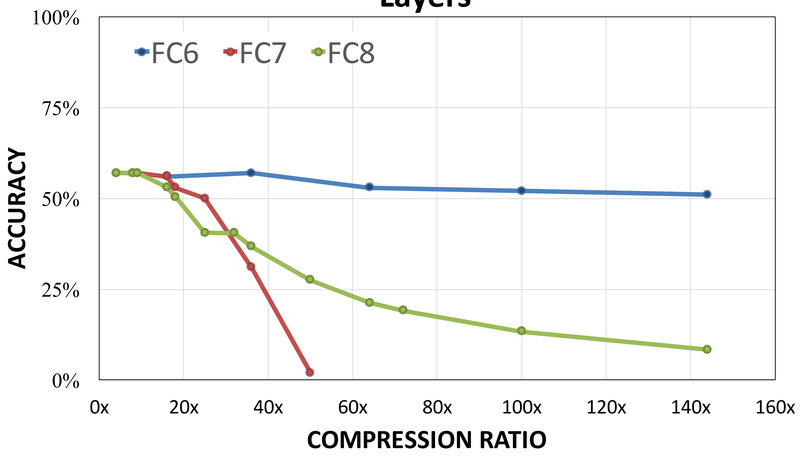
ASTC Re-trained weights (AlexNet)



Sensitivity to ASTC compression

Effect of compression on the three fully-connected layers of AlexNet: FC6 (9216x4096) > FC7 (4096x4096) > FC8 (4096x1000). Thus, different layers can be compressed with different compression ratios to obtain a better accuracy-compression tradeoff.

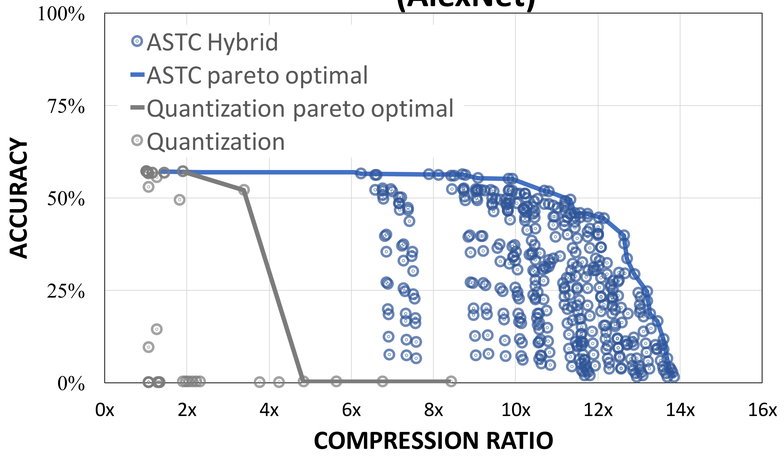
Effect of ASTC Compression on AlexNet Layers



Pareto Analysis

When using a different compression parameter for each layer, the number of points can be large.
Below, we show the Pareto optimal curve for the accuracy vs compression ratio tradeoff.

Pareto Analysis: Accuracy vs Compression (AlexNet)



†Hardik Sharma
hsharma@gatech.edu

Tom Olson
tom.olson@arm.com

Alex Chalfin
alex.chalfin@arm.com

† Georgia Institute of Technology
ARM, Inc.