Summary of Paper: Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

Muthu Chidambaram

Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/

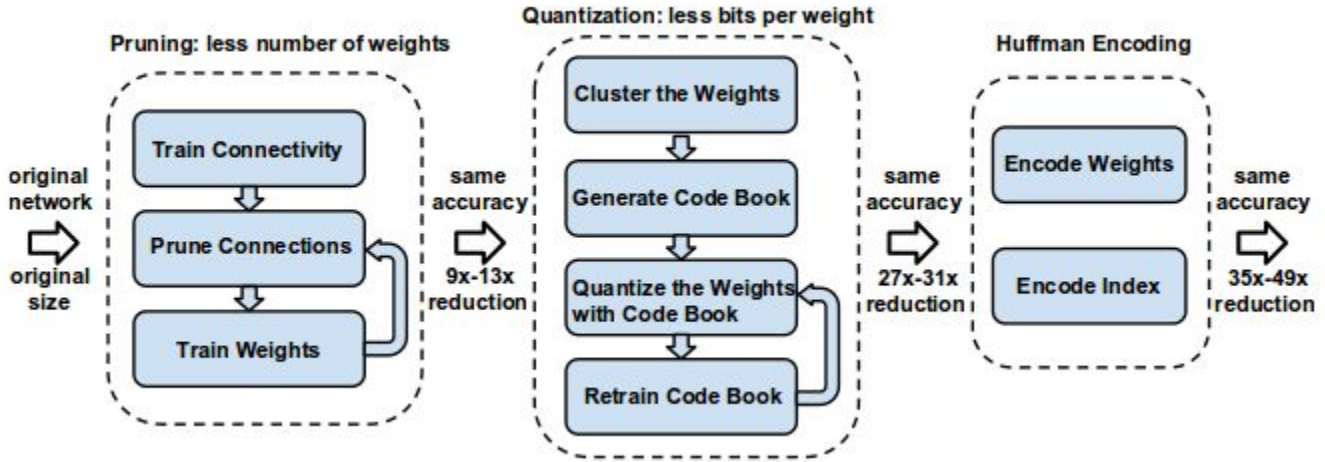
Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

- Authors: Song Han, Huizi Mao, William J. Dally
- Goal: reduce computational and memory intensiveness of neural networks through pruning, quantization, and encoding
- Has immediate applications to mobile development using neural nets

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

- Goal is to reduce storage and energy of neural network models so as to employ them on mobile devices
- Key insight is that pruning and quantization can be done without interfering with one another
- Network pruning is done by first learning all connections, then removing weights below a certain threshold

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)



Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

- Uses compressed sparse row (CSR) + storing index difference instead of absolute position to store pruned weights
- Reduce number of connections needed to be stored by sharing weights across connections
  - Weights quantized into bins

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

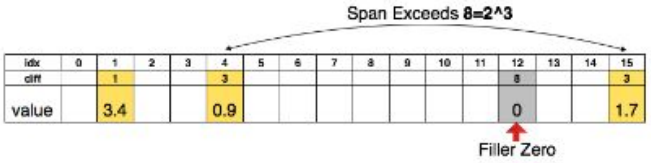


Figure 2: Representing the matrix sparsity with relative index. Padding filler zero to prevent overflow.

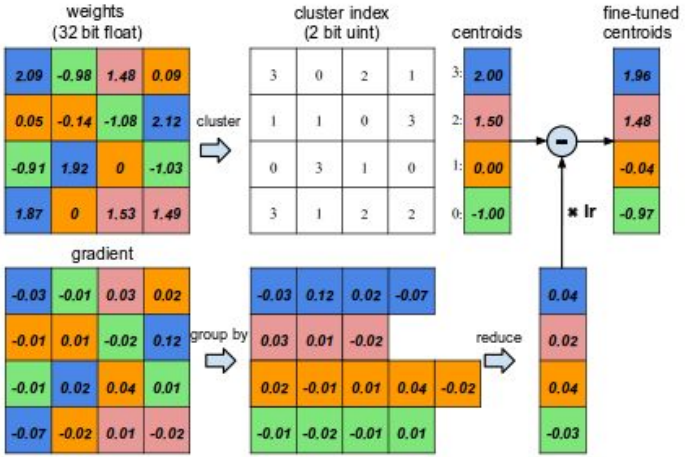


Figure 3: Weight sharing by scalar quantization (top) and centroids fine-tuning (bottom).

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

- Use k-means to determine which weights are shared for each layer of a network
  - 3 different centroid initialization methods: forgy (random), density-based, and linear
- During backprop, the gradient for each shared weight is calculated and used to update the shared weight
- Huffman encode quantized weights to further save space

$$\frac{\partial \mathcal{L}}{\partial C_k} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial W_{ij}} \frac{\partial W_{ij}}{\partial C_k} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial W_{ij}} \mathbb{1}(I_{ij} = k)$$

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

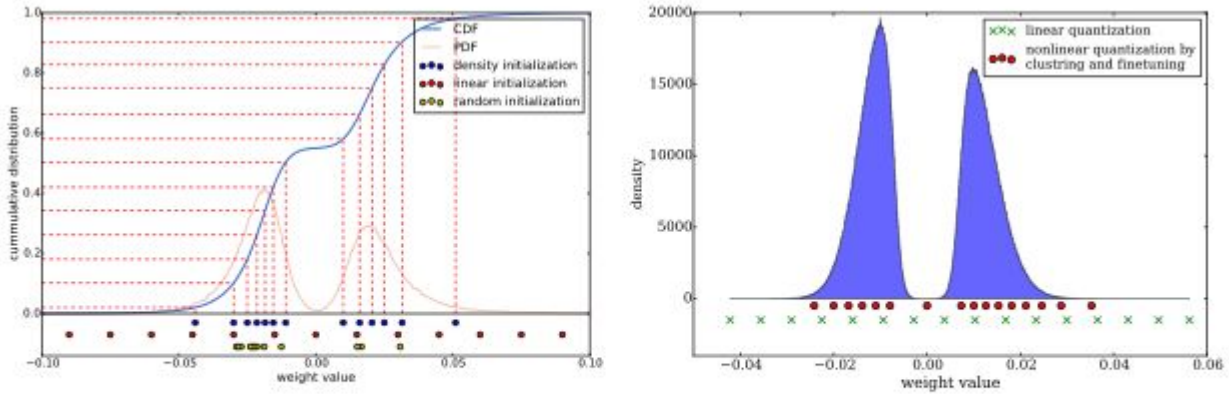


Figure 4: Left: Three different methods for centroids initialization. Right: Distribution of weights (blue) and distribution of codebook before (green cross) and after fine-tuning (red dot).

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

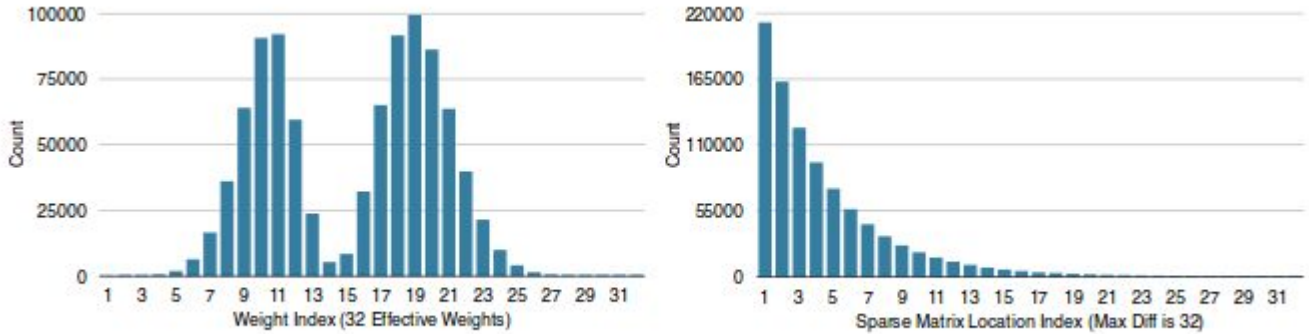


Figure 5: Distribution for weight (Left) and index (Right). The distribution is biased.

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

Table 1: The compression pipeline can save 35× to 49× parameter storage with no loss of accuracy.

Network	Top-1 Error	Top-5 Error	Parameters	Compress Rate
LeNet-300-100 Ref	1.64%	-	1070 KB	
LeNet-300-100 Compressed	1.58%	-	**27 KB**	**40×**
LeNet-5 Ref	0.80%	-	1720 KB	
LeNet-5 Compressed	0.74%	-	**44 KB**	**39×**
AlexNet Ref	42.78%	19.73%	240 MB	
AlexNet Compressed	42.78%	19.70%	**6.9 MB**	**35×**
VGG-16 Ref	31.50%	11.32%	552 MB	
VGG-16 Compressed	31.17%	10.91%	**11.3 MB**	**49×**

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

– – –

Table 2: Compression statistics for LeNet-300-100. P: pruning, Q:quantization, H:Huffman coding.

Layer	#Weights	Weights% (P)	Weight bits (P+Q)	Weight bits (P+Q+H)	Index bits (P+Q)	Index bits (P+Q+H)	Compress rate (P+Q)	Compress rate (P+Q+H)
ip1	235K	8%	6	4.4	5	3.7	3.1%	2.32%
ip2	30K	9%	6	4.4	5	4.3	3.8%	3.04%
ip3	1K	26%	6	4.3	5	3.2	15.7%	12.70%
Total	266K	8%(12×)	6	5.1	5	3.7	3.1% (**32**×)	2.49% (**40**×)

Table 3: Compression statistics for LeNet-5. P: pruning, Q:quantization, H:Huffman coding.

Layer	#Weights	Weights% (P)	Weight bits (P+Q)	Weight bits (P+Q+H)	Index bits (P+Q)	Index bits (P+Q+H)	Compress rate (P+Q)	Compress rate (P+Q+H)
conv1	0.5K	66%	8	7.2	5	1.5	78.5%	67.45%
conv2	25K	12%	8	7.2	5	3.9	6.0%	5.28%
ip1	400K	8%	5	4.5	5	4.5	2.7%	2.45%
ip2	5K	19%	5	5.2	5	3.7	6.9%	6.13%
Total	431K	8%(12×)	5.3	4.1	5	4.4	3.05% (**33**×)	2.55% (**39**×)

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

– – –

Table 4: Compression statistics for AlexNet. P: pruning, Q: quantization, H:Huffman coding.

Layer	#Weights	Weights% (P)	Weight bits (P+Q)	Weight bits (P+Q+H)	Index bits (P+Q)	Index bits (P+Q+H)	Compress rate (P+Q)	Compress rate (P+Q+H)
conv1	35K	84%	8	6.3	4	1.2	32.6%	20.53%
conv2	307K	38%	8	5.5	4	2.3	14.5%	9.43%
conv3	885K	35%	8	5.1	4	2.6	13.1%	8.44%
conv4	663K	37%	8	5.2	4	2.5	14.1%	9.11%
conv5	442K	37%	8	5.6	4	2.5	14.0%	9.43%
fc6	38M	9%	5	3.9	4	3.2	3.0%	2.39%
fc7	17M	9%	5	3.6	4	3.7	3.0%	2.46%
fc8	4M	25%	5	4	4	3.2	7.3%	5.85%
Total	61M	11%(9×)	5.4	4	4	3.2	3.7% (**27**×)	2.88% (**35**×)

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

– – –

Table 5: Compression statistics for VGG-16. P: pruning, Q:quantization, H:Huffman coding.

Layer	#Weights	Weights% (P)	Weigh bits (P+Q)	Weight bits (P+Q+H)	Index bits (P+Q)	Index bits (P+Q+H)	Compress rate (P+Q)	Compress rate (P+Q+H)
conv1_1	2K	58%	8	6.8	5	1.7	40.0%	29.97%
conv1_2	37K	22%	8	6.5	5	2.6	9.8%	6.99%
conv2_1	74K	34%	8	5.6	5	2.4	14.3%	8.91%
conv2_2	148K	36%	8	5.9	5	2.3	14.7%	9.31%
conv3_1	295K	53%	8	4.8	5	1.8	21.7%	11.15%
conv3_2	590K	24%	8	4.6	5	2.9	9.7%	5.67%
conv3_3	590K	42%	8	4.6	5	2.2	17.0%	8.96%
conv4_1	1M	32%	8	4.6	5	2.6	13.1%	7.29%
conv4_2	2M	27%	8	4.2	5	2.9	10.9%	5.93%
conv4_3	2M	34%	8	4.4	5	2.5	14.0%	7.47%
conv5_1	2M	35%	8	4.7	5	2.5	14.3%	8.00%
conv5_2	2M	29%	8	4.6	5	2.7	11.7%	6.52%
conv5_3	2M	36%	8	4.6	5	2.3	14.8%	7.79%
fc6	103M	4%	5	3.6	5	3.5	1.6%	1.10%
fc7	17M	4%	5	4	5	4.3	1.5%	1.25%
fc8	4M	23%	5	4	5	3.4	7.1%	5.24%
Total	138M	7.5%(13×)	6.4	4.1	5	3.1	3.2% (**31**×)	2.05% (**49**×)

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

- Quantization works well with pruning, because the number of centroids is fixed while there are fewer weights

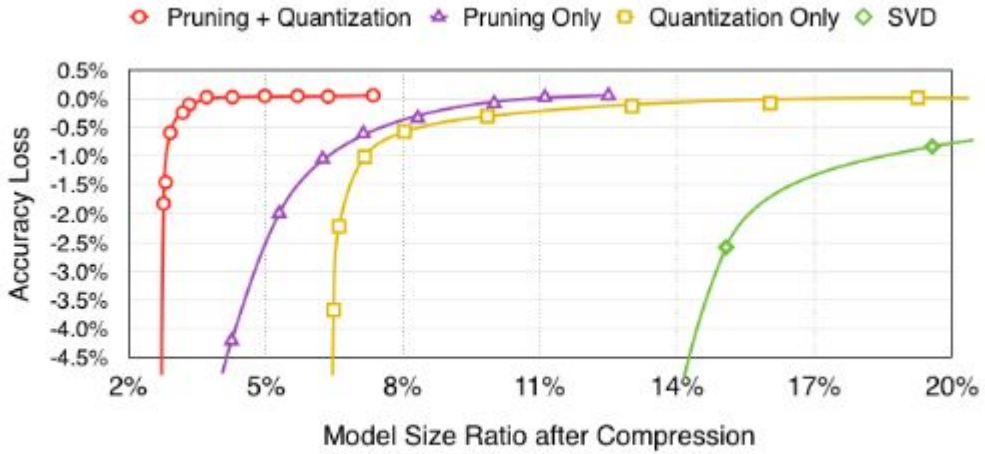


Figure 6: Accuracy v.s. compression rate under different compression methods. Pruning and quantization works best when combined.

Deep Compression: Compressing Deep Neural Networks (ICLR 2016)

- Conclusion: deep compression significantly compresses networks without affecting accuracy
- Future work: quantized network with weight sharing requires either customized gpu kernels or specialized asic architecture