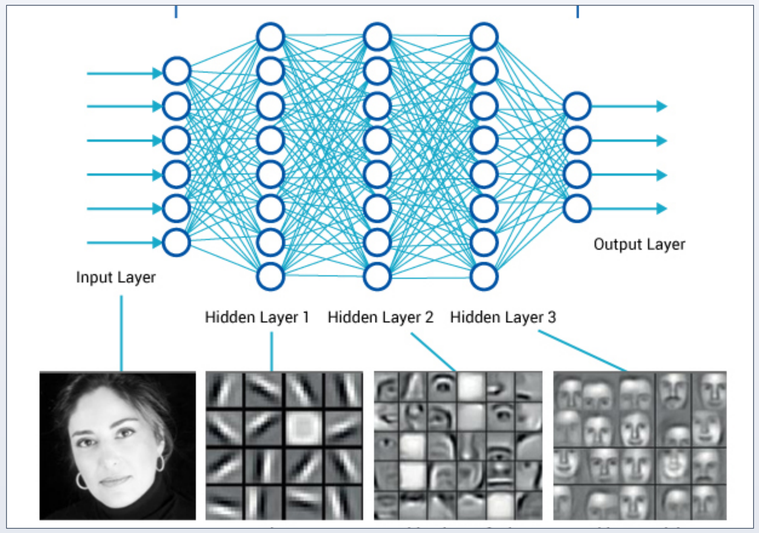
Compression of Deep Neural Networks

Tuguldur Baigalmaa

Earlham College Computer Science Department

INTRODUCTION



Deep Neural Networks (DNNs) are a set of algorithms, inspired by the inner workings of the human mind. Recent breakthroughs have made DNNs the industry standard for solving problems in computer vision, speech recognition and natural language processing. Due to the need to deploy DNN models on a variety of platforms, such as mobile devices and embedded systems with limited hardware capabilities and resources, compressed and scalable representations of the network parameters are in urgent need not only to reduce storage and network bandwidth requirements, but also run inferences of the model more efficiently. We studied the impact of compression techniques on some of the most popular image recognition DNNs.

OBJECTIVES

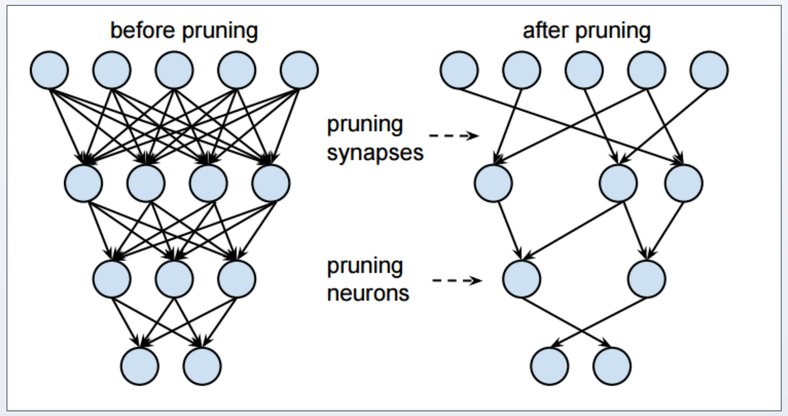
The overarching objective of the research is to survey a variety of DNN compression techniques proposed by the research community and study their impact on model inference and network parameter size. Another objective of the research is to study the feasibility of deploying compressed, sophisticated DNN models to mobile devices for better privacy, less network bandwidth and real time processing.

COMPRESSION TECHNIQUES

Han et al. have proposed a compression pipeline in their paper "Deep Compression" [1], which involves network pruning, quantization and Huffman coding.

NETWORK PRUNING

Network trimming/pruning has been well-studied to compress Convolutional Neural Networks (CNNs). Recently, Han et al. detailed an approach to prune CNN models with no loss of accuracy [2]. Their approach entails removing connections below a certain threshold after an initial training phase. The resulting sparse network is then retrained to make appropriate adjustments to the weights of the remaining connections.



The network pruning process mimics the mammalian brain, where the synapses that were created during childhood are gradually pruned by reducing the connections that are not used as much [3].

QUANTIZATION & WEIGHT SHARING

Quantization reduces the number of bits required store each weight by having multiple connections share the same weight. Quantizing reduces the computational resources needed to run inference calculations by operating on a lower-precision arithmetic, weights being represented in 8-bits. ML libraries, such as TensorFlow, have already built automatic quantization routines since it provides an easy performance improvement.

HUFFMAN CODING

Huffman coding is a common lossless data compression technique of using codewords to encode source symbols. The method can be used to encode both weights and indexes. Huffman coding non-uniformly distributed values results in additional 20%-30% storage reduction.

RESULT

Pruning state of the art image recognition networks, such as AlexNet and VGG-16, reduced the number of parameters by 9X and 13X respectively.

The table [1] below specifies the compression statistics for applying the compression pipeline on AlexNet.

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**×)

Choi et al. have improved upon the compression pipeline by using Uniform Quantization and Hessian-weighted k-means, reaching 40X-51X on AlexNet and LeNet respectively, compared to 35X-39X detailed in Han et al [4].

			Accuracy %	Compression ratio
LeNet5	Original model		99.25	-
	Pruned model		99.27	10.13
	Pruning + Quantization all layers + Huffman coding	k-means	99.27	44.58
		Hessian-weighted k-means	99.27	47.16
		Uniform quantization	99.28	51.25
		Iterative ECSQ	99.27	49.01
	Deep compression (Han et al., 2015a)		99.26	39.00
ResNet	Original model		92.58	-
	Pruned model		92.58	4.52
	Pruning + Quantization all layers + Huffman coding	k-means	92.64	18.25
		Hessian-weighted k-means	92.67	20.51
		Uniform quantization	92.68	22.17
		Iterative ECSQ	92.73	21.01
AlexNet	Original model		57.16	-
	Pruned model		56.00	7.91
	Pruning + Quantization all layers + Huffman coding	k-means	56.12	30.53
		Hessian-weighted k-means	56.04	33.71
		Uniform quantization	56.20	40.65
	Deep compression (Han et al., 2015a)		57.22	35.00

The table [4] above displays the statistics for using alternative compression approaches. The compression rate ranges from 18X to 51X while preserving the original accuracy of the model.

CONCLUSION

The compression techniques for DNNs are producing outstanding results and deploying DNN models into mobile devices is becoming more feasible. However, the research is still in early stages and most of the DNNs that were experimented on were CNNs used for image recognition tasks. However, the same techniques, such as network pruning, quantization, weight sharing and huffman coding, will work for a variety of DNNs since they all share the same underlying structural components. Thanks to the reduction in model size and increase in efficiency of calculating inferences, more and more sophisticated models will be deployed on the mobile platform as the underlying software infrastructure becomes available.

REFERENCES

Figure 1. Amax. 2016. Retrieved Dec. 1th, 2016, from www.amax.com/blog/wp-content/uploads/2015/12/blog_deeplearning3.jpg.

Figure 2. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding". Retrieved Dec. 1th, 2016, from https://arxiv.org/abs/1510.00149

[1] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding" *arXiv:1510.00149 [cs]*, Oct. 2015.

[2] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both Weights and Connections for Efficient Neural Networks," *arXiv:1506.02626 [cs]*, Jun. 2015.

[3] JP Rauschecker. Neuronal mechanisms of developmental plasticity in the cat's visual system. Human neurobiology, 3(2):109–114, 1983.

[4] Y. Choi, M. El-Khamy, and J. Lee, "Towards the Limit of Network Quantization," *arXiv:1612.01543 [cs]*, Dec. 2016.

ACKNOWLEDGEMENTS

- Earlham College Computer Science Department
- David Michael Barbella (Advisor)
- Xunfei Jiang (CS Faculty)
- Researchers studying compression of deep neural networks