Paper Reading Report

Hu Jiyuan

September 17, 2018

1 Requirements

Draw an outline for the paper and summarize the main idea of each part.

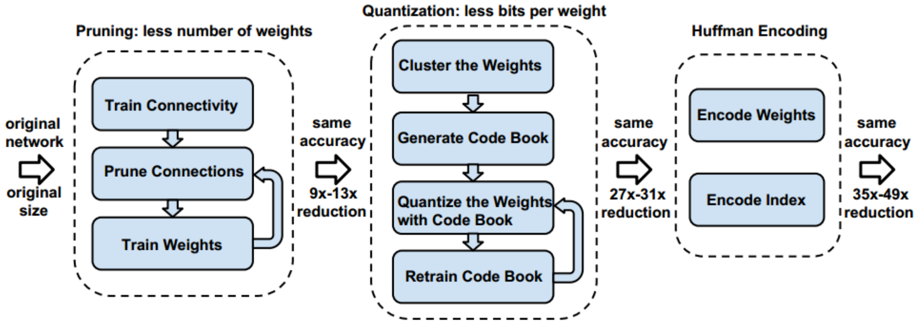
Quote sentences typical of a part that you can use in your own paper.

Quote sentences you find difficult to understand.

Record the bibliographic information of the paper (1) by following its documentation style and (2) in the Chicago style.

2 Report

2.1 Outline & main idea



ABSTRACT answers the following four questions:
What problems have people met?
What's your solution?
How does the solution work?
How well does the solution work in comparison of existing solutions?

Part 1: INTRODUCTION first introduces the background and need of the neural networks compression solutions: reduce the storage and energy required to run inference on large networks. Then briefly describes their solutions in three parts: pruning, trained quantization and Huffman coding.

Part 2: NETWORK PRUNING shows how to prune the network. Neural networks have many connections, they prune the small-weight connections.

Part 3: TRAINED QUANTIZATION AND WEIGHT SHARING continues to compress the network by reducing the number of bits required to represent each weight, which is what called quantization. They use k-means clustering to find the weight that can be shared.

Part 4: HUFFMAN CODING is used in the purpose of encoding (or compression) the quantized weights and sparse matrix index generated by the former step.

Part 5: EXPERIMENTS are taken to show the performance of their solutions by comparing the network parameters and accuracy of several networks before and after their solutions are taken.

Part 6: DISCUSSIONS discuss the relationship between performance and configuration.

Part 7: RELATED WORK is the extension of the background description in Part 1: introduction. It mainly tells us the neural networks are over-parametrized and introduce the typical methods taken to solve the problem by other people.

Part 8: FUTURE WORK denotes what they are going to do next: build software to complete the benchmarked test they does not finish because of the existing software does not support their solution. What's more, to build hardware (ASIC chip) to achieve their solutions.

Part 9: CONCLUSION concludes their steps, key performance and application scene of their solutions.

REFERENCES are in alphabetical order.

2.2 Typical sentences

To address this limitation, we introduce something, some introduction.

Examples: To address this limitation, we introduce deep compression, a three stage pipeline: pruning, trained quantization and Huffman coding, that work together to reduce the storage requirement of neural networks by 35 to 49 without affecting their accuracy.

To achieve this goal, we present deep compression: a three stage pipeline (Figure 1) to reduce the storage required by neural network in a manner that preserves the original accuracy.

Figure xxx is illustrated in Figure x.

Examples: Weight sharing is illustrated in Figure 3.

2.3 Difficult sentences

While the *pruned* network has been benchmarked on various hardware, the *quantized* network with weight sharing has not, because off-the-shelf cuSPARSE or MKL SPBLAS library does not support indirect matrix entry lookup, nor is the relative index in CSC or CSR format supported. So the full advantage of Deep Compression that fit the model in cache is not fully unveiled.

2.4 Bibliographic information

Han, Song, Mao, Huizi and Dally, William J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016

Han, Song, Mao, Huizi and Dally, William J. “Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding.” International Conference on Learning Representations (2016). https://arxiv.org/abs/1510.00149

Google scholar: Han, Song, Huizi Mao, and William J. Dally. “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding.” *arXiv preprint arXiv:1510.00149* (2015).