PAPER READING REPORT

Hu Jiyuan [*]
School of Electronics and Information Technology
Sun Yat-sen University
Guangzhou, China
hujy23@mail2.sysu.edu.cn

ABSTRACT

[1] 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.

1 OUTLINE & MAIN IDEA

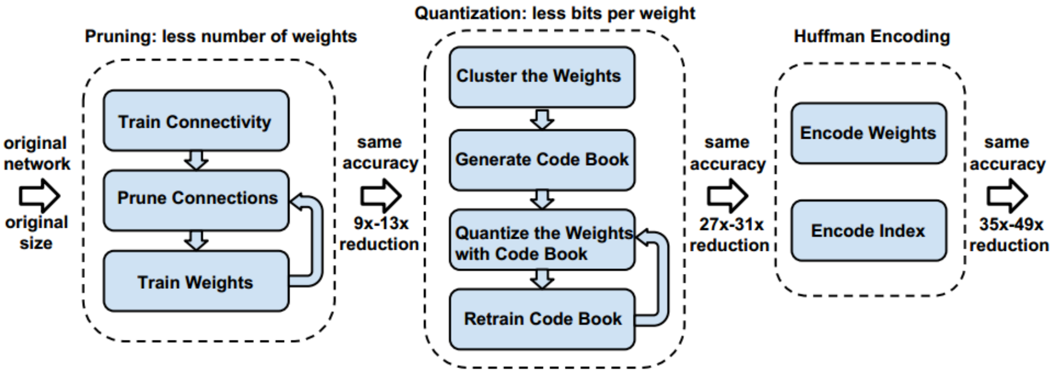

Figure 1: Main content of the paper I read

1.1 ABSTRACT

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?

1.2 INTRODUCTION

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.

[*] Student ID: 18215258, Class ID: 11

[1] I originally want to use "REQUIREMENTS" here but this report is organized by the same LaTeX template with the paper I read, I can't change the "ABSTRACT" and the header. "Under review as a conference paper at ICLR 2016"

1.3 NETWORK PRUNING

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

1.4 TRAINED QUANTIZATION AND WEIGHT SHARING

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.

1.5 HUFFMAN CODING

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

1.6 EXPERIMENTS

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.

1.7 DISCUSSIONS

DISCUSSIONS discuss the relationship between performance and configuration.

1.8 RELATED WORD

RELATED WORD is the extension of the background description in section 1.2. It mainly tells us the neural networks are over-parametrized and introduce the typical methods taken to solve the problem by other people.

1.9 FETURE WORK

FETURE 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.

1.10 CONCLUSION

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

1.11 REFERENCES

REFERENCES are in alphabetical order.

2 TYPICAL SENTENCES

2.1 TO ADDRESS

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.

2.2 FIGURE

xxx is illustrated in Figure x.

Examples: The main content is illustrated in Figure 1.

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.

4 BIBLIOGRAPHIC INFORMATION

4.1 WRITTEN BY MYSELF

4.1.1 STYLE USED BY THE AUTHOR

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

4.1.2 CHICAGO STYLE

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

4.2 STANDARD ANSWER

4.2.1 STYLE USED BY THE AUTHOR

This is in-text citation, see Han et al. (2015).

And the following is bibliography:

REFERENCES

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

Generated by the ICLR LaTeX template itself.

4.2.2 CHICAGO STYLE

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).

Generated by Google Scholar.