ECE 561 Literature Survey: Optimization Techniques to Reduce Memory Footprints and Inference Time

Anisha Aswani (anisha.aswani@colostate.edu)

7th September 2019

1 Introduction

This literature survey is based mainly on different optimization techniques to reduce memory footprints and inference time of a neural network. The available computational resources often has constraints while training a large scaled deep neural networks. Using neural network algorithms on conventional general purpose digital hardware has been found highly inefficient due to the massive amount of multiply-accumulate operations required to compute the weighted sums of the neurons[5]. One of the main issue of machine learning is the inference efficiency and with all these restriction it is difficult to deploy neural networks on embedded systems which have limited hardware resources. The different methods that will be studied during this literature survey is the use of quantization and binarization to reduce the inference time.

2 Scope

The scope of this literature survey is to get in depth knowledge of different ways to reduce inference time of a neural network. Along the way, the application of quantization and binarization will be observed on a neural network system. In the big picture, this literature survey will help in better implementation of machine learning algorithms on low power devices.

3 Progress Report

A thorough research has been done on the use of quantization to reduce memory footprint. Works by Gupta et al.[2], Shuang Wu et al.[3], Penghang Yin et al.[4], Matthieu Courbariaux et al.[5], Jacob, Benoit, et al.[6], Courbariaux, Matthieu et al.[7], Han, Song et al.[9], Raghuraman Krishnamoorthi[10] have been surveyed. Future work of this literature survey will be based on the use of BinaryNet and TernaryNet on more efficient use of memory on low power

devices. Further reading will be done on papers referenced [11] to [30] to study more about different ways to reduce memory usage while using neural networks. Different algorithms will be studied along with their applications and compared for the final report.

References

[1] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, Yoshua Bengio, "Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or -1". http://papers.nips.cc/paper/6573-binarized-neural-networks.pdf

[2] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, Pritish Narayanan, " Deep Learning with Limited Numerical Precision" arXiv: 1502.02551 https://arxiv.org/pdf/1502.02551.pdf

[3] Shuang Wu, Guoqi Li, Feng Chen, Luping Shi, " Training and inference with integers in deep neural networks", ICLR 2018. https://openreview.net/pdf?id=HJGXzmspb

[4] Blended Coarse Gradient Descent for Full Quantization of Deep Neural Networks. https://arxiv.org/pdf/1808.05240.pdf

[5] Quantized Neural Networks:Training Neural Networks with Low Precision weights and activation. https://arxiv.org/pdf/1609.07061.pdf

[6] Jacob, Benoit, et al. "Quantization and training of neural networks for efficient integer-arithmetic-only inference." arXiv preprint arXiv:1712.05877 (2017).
http://openaccess.thecvf.com/content_cvpr_2018/papers/Jacob_Quantization_and_Training_CVPR_2018_paper.pdf

[7] Courbariaux, Matthieu, Yoshua Bengio, and Jean-Pierre David. "Training deep neural networks with low precision multiplications."arXiv preprint arXiv:1412.7024 (2014). https://arxiv.org/pdf/1412.7024.pdf

[8] Zhu,Chenzhuo, etal." Trained ternary quantization." arXiv preprint arXiv:1612.01064(2016) https://arxiv.org/pdf/1612.01064.pdf

[9] 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). https://arxiv.org/pdf/1510.00149.pdf

[10] Raghuraman Krishnamoorthi."Quantizing deep convolutional networks for efficient inference." arXiv: 1806.08342 https://arxiv.org/pdf/1806.08342.pdf

[11] Yunchao Gong, Liu Liu, Lubomir Bourdev."Compressing Deep Convolutional Network using Vecotr Quantization." arXiv: 1412.611 https://arxiv.org/pdf/1412.6115.pdf

[12] Sergey Ioffe, Christian Szegedy."Batch Normalization: Accelerating Deep Network Training by Reducing Inernal Cova." arXiv: 1502.03167 https://arxiv.org/pdf/1502.03167.pdf

[13] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, Jian Cheng."Quantized Convolutional Neural Networks for Mobile Devices." arXiv: 1512.06473 https://arxiv.org/pdf/1512.06473.pdf

[14] Minje Kim, Paris Smaragdis."Bitwise Neural Networks." arXiv: 1601.0607 https://arxiv.org/pdf/1601.0607.pdf

[15] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, William J. Dally."EIE: Efficient Inference Engine on Compressed Deep Neural Network." arXiv: 1602.01528 https://arxiv.org/pdf/1602.01528.pdf

[16] Mohammad Rastegariy, Vicente Ordonezy, Joseph Redmon, Ali Farhadiy."XNOR-Net: ImageNet Classification Using Binary." Convolutional Neural Networks." arXiv: 1603.05279 https://arxiv.org/pdf/1603.05279.pdf

[17] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, Yuheng Zou, "DOREFA-NET: TRAINING LOW BITWIDTH CONVOLUTIONAL NEURAL NETWORKS WITH LOW BITWIDTH GRADIENTS." arXiv: :1606.06160 https://arxiv.org/pdf/:1606.06160.pdf

[18] Lu Hou, Quanming Yao, James T. Kwok."LOSS-AWARE BINARIZATION OF DEEP NETWORKS." arXiv:1611.01600 https://arxiv.org/pdf/1611.01600.pdf

[19] Chenzhuo Zhu et al."TRAINED TERNARY QUANTIZATION." arXiv: :1612.01064 https://arxiv.org/pdf/:1612.01064.pdf

[20] Yu Cheng, Duo Wang, Pan Zhou, Tao Zhang, ."A Survey of Model Compression and Acceleration for Deep Neural Networks." arXiv: 1710.09282 https://arxiv.org/pdf/1710.09282.pdf

[21] Bert Moons, Koen Goetschalckx, Nick Van Berckelaer, and Marian Verhelst, "Minimum Energy Quantized Neural Networks." arXiv:1711.00215 https://arxiv.org/pdf/1711.00215.pdf

[22] Song Han et al."Learning both Weights and Connections for Efficient Neural Networks."

[23] Alexander Novikov, Dmitry Podoprikhin1 Anton Osokin, Dmitry Vetrov, "Tensorizing Neural Networks."

[24] Dongyoung Kim et al. "A Novel Zero Weight/Activation-Aware Hardware Architecture of Convolutional Neural Network." IEEE,2017: 978-3-9815370-8-6/17

[25] Tianshi Chen et al."DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning."

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database."

[27] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[28] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In Advances in Neural Information Processing Systems, pp.3123–3131, 2015.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," arXiv:1409.4842, 2014.

[30] A. Shafiee and et al., "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," ISCA, 2016.