Only Integer Multiplication Neural Networks for On Device AI

You Jun Kim, Choong Seon Hong
Department of Computer Science and Engineering
Kyung Hee University, 17104
Republic of Korea
{yj4889, cshong}@khu.ac.kr

***Abstract*—Recently, training aware quantization using learnable quantization parameters is emerging. The above method is drawing attention because the accuracy of the 4bit quantization model is almost similar to that of the full-precision model. In addition, when the batch normalization layer, which plays an important role in the convolutional neural network, is folded with the convolution layer, it is possible to increase the inference speed because batch-related operations are unnecessary in model inference. This paper proposes a method of folding the convolution layer and the batch normalization layer into one module, and quantizing the weight and activation of the folded module with learnable quantization parameters. It was confirmed that the result of the proposed method was higher than Asymmetric Uniform quantization(IAO), which is a beta version of library provided by Pytorch, as VGG16: +0.2, ResNet18: +0.07, MobileNetv2: +3.32. In addition, it was confirmed that MovileNetv2 was impossible with the Pytorch library. The proposed method can be used for IoT (Internet of Things), Edge, etc. that require real-time inference.**

Keywords—Learnable Quantization Parameters, folding, CNN

I. Introduction

In this paper, we propose a method of training the learnable quantization parameter, the weight and bias of the convolution (linear) layer, and the weight and bias of the batch normalization layer in backpropagation at the same time.

II. Method

A. Batch Normalization folding

Batch Normalization[1] is an essential technique in many AI models by adjusting the data distribution to enable stable learning. In [2], batch normalization layer can be folded with convolution(linear) layer to become one layer.

B. Quantization aware training-based batch folding and learnable quantizaion parameters

We propose a method of obtaining high accuracy through optimal quantization using batch folding and Learnable quantization parameters that enable faster inference. Figure 1 shows the training, fine-tuning, and inference flow of the proposed method. Training has the same flow as general model training (Conv→Batch→ReLu). The fine tuning process freezes the moving average and moving variance of the batch normalization layer[3] and The weight(W) and bias(b) of the convolution layer, the weight(γ) and bias(B) of the batch normalization, and learnable quantization parameters(s_w, s_a) are trained to be optimized for fake quantization. We used [4] quantization method(LSQ).

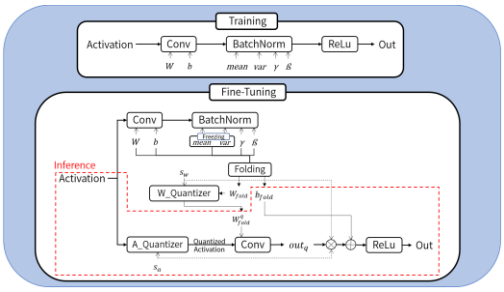

Figure 1. Training, Fine-Tuning, and Inference flow of the proposed method.

III. Result

We used cifar10 dataset and experimented with VGG16[5], ResNet18[6], MobileNetv2[7].

Table 1. Proposed Method vs IAO(integer-arithmetic-only inference)[3]

Method(Folding Batch norm + Quantization)	VGG16	ResNet18	MovileNetv2
32bits float	93.24	93.87	92.60
Learnable Quantization Parameter(4bits)	93.35	93.87	92.40
Asymmetric Uniform Quantization(4bits)	93.18	93.8	88.08

IV. Conclusion

The proposed method uses quantization aware training-based batch folding and learnable quantization parameters, eliminating the need for batch-related operations and obtaining an optimal quantization model. This allows the model in inference faster and has high accuracy at 4bits quantization.

References

[1] IOFFE, Sergey; SZEGEDY, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[2] LI, Rundong, et al. Fully quantized network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. p. 2810-2819.

[3] JACOB, Benoit, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 2704-2713.

[4] ESSER, Steven K., et al. Learned step size quantization. arXiv preprint arXiv:1902.08153, 2019.

[5] SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[6] HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[7] SANDLER, Mark, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 4510-4520