https://doi.org/10.2352/EI.2022.34.7.ISS-183

On quantization of convolutional neural networks for image restoration

Youngil Seo, Irina Kim, Jungguk Lee, Wooseok Choi, Seongwook Song; Samsung Electronics Ltd; Hwaseong/Korea

Abstract

Recently, commercial vision sensors hit the mobile market. To achieve that, computer vision networks had to be quantized. However, this topic was not studied well for Image Signal processor (ISP) challenging image restoration tasks, being crucially important for hardware implementation, as well as for deployment on hardware accelerators, e.g. Neural Processors Units (NPU).

In this paper, we studied the effect of the quantization of deep learning network on image quality. We tried various quantization on raw RGBW image demosaicing. Experimental results show that 10 bit weight quantization can sustain image quality at the similar level with floating-point network. 8 bit quantized network shows slight degradation in objective image quality and mild visual artifacts.

Although network weight's bit-depth can be significantly reduced for computer vision tasks, our finding shows that it is not true for raw image restoration tasks: at least 10 bit weights are required to provide sufficient image quality. However, we can save some memory on feature maps bit-depth. We can conclude that network bit depth is critical for raw image restoration.

Introduction

Deep neural networks (DNNs) have become the state-of-the-art in the computer vision and sequence modeling problems like image classification, object detection, speech recognition. However, they usually comes with high cost computation and memory costs from a huge amount of parameters. For example, Reference [1] came up with a deep convolutional neural network consisting of 61 million parameters and won the ImageNet competition in 2012. It is followed by deeper neural networks with even larger numbers of parameters, e.g., Reference [2]. So it makes the deployment on the mobile platforms that have limited power and computation resources challenging.

Parameters of a trained neural network commonly exhibit high degrees of redundancy [3] which implies an over-parametrization of the network. Network compression methods implicitly or explicitly aim at the systematic reduction of redundancy in neural network models while at the same time retaining a high level of task accuracy [4]. Many compression methods perform some form of pruning or quantization. Pruning is the removal of irrelevant units (weights, neurons or convolutional filters)[5]. Quantization is the reduction of the bit-precision of weights, activations or even gradients, which is particularly desirable from a hardware perspective[6]

Network quantization for vision applications like classification, image segmentation and object detection has drawn great attention of researchers [1] [2] [7] [8] [9]. Approaches for low-bit quantization of neural networks have been made for these applications. There are binary weight networks [10] [11] and ternary networks [12] [13] [14]. But owing to requirement of high bit-depth and high resolution there are no prior art on quantization of image restoration problems like demosaicing, super resolution and deblurring, etc.

In this paper, we studied the effect of the quantization of deep learning network on image quality. We tried to test various quantization bit-depth that is not supported in the conventional AI platform like Tensorflow or Pytorch and AI hardware like NPU, DSP or GPU. We tried to find what is optimal bit-depth without image quality degradation. We could also find the best suitable bit-depth for custom AI hardware.

Releated works

In this work we focus mostly on demosaicing and quantization algorithms, so we will brifly review related works.

Demosaicing of bayer color filter array has been extremely studied. [15], [16].There are various conventional approches, such as color difference based interpolation [17], [18], frequency domain filtering [19], [20], [21], and reconstruction methods [22], [23]. But for new other patterns, other effort like hand-crafted algorithms should be applied to solve it. So there is also universal approach [24].

Deep learning approaches to demosaicing has been applied [25], [26], [27], [28]. Previously, many researches focused on the bayer CFA demosaicing, but there are researches on Quad bayer pattern and Nona pattern demosaicing also [29], [30]. Deep learning methods have better image quality in complex CFA pattern demosaicing although they require high computation cost.

Especially we focus on RGBW CFA and its demosaicing. There are conventional algorithms like [33], [34] and deep learning approaches like [35]. Here our approach is related to deep learning RGBW demosaicing.

To deploy deep network on mobile platform, quantization is needed usually. There are two types of quantization methods. It is often desirable to reduce the model size by quantizing weights and activations post-training, without the need to re-train/fine-tune the model. These methods, commonly referred to as post-training quantization, are simple to use and allow for quantization with limited data [31].Quantization-aware training simulates quantization during training so that the quantization parameters can be learned together with the model using training data [32].

Problem statement

Demosaicing can be done for any CFA pattern, but there are image quality degradations when using non-standard CFA patterns, such as RGBW. To enhance image quality for RGBW CFA demosaicing, we aim to use deep neural network as shown in Fig. 1. Deplyment on a mobile platform requires quantization of network. In this work we studied to find how many bits are sufficient for the quantized network without image quality degradation.

dstributedattention

Proposed method

In this work we developed our own network called DePhaseNet. Its features are multi-level network with multi-phase inputs to adopt various phase schemes and correlations, distributed attention modules for layer-wise proportional sum and frequency loss training scheme. In Fig. 2 PE means phase extractor that extracts various combination of phases and subnet consists of residual blocks and attention blocks. Fig. 3 shows distributed attention module.

The network architecture shows outperforming image quality and artifact reduction compared to previous methods. But it comes at cost 1.9MMAC/1 pixel. Deployment of this network on device needs quantization.

In this work we used post training quantization to see the direct effect of quantization on image quality. There are two quantizations in the network showed in Fig. 4, one is weight quantization and the other is feature map (activation) quantization. However, we did weight quantization only putting activations floating point in this study, to see how weight bit depth affects image quality. We tested quantization with various bits.

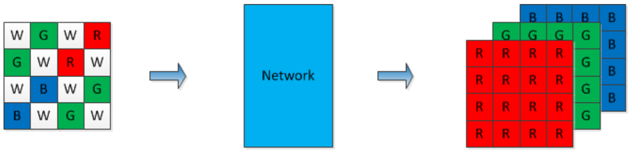

Figure 1. RGBW demosaicing with deep learning network

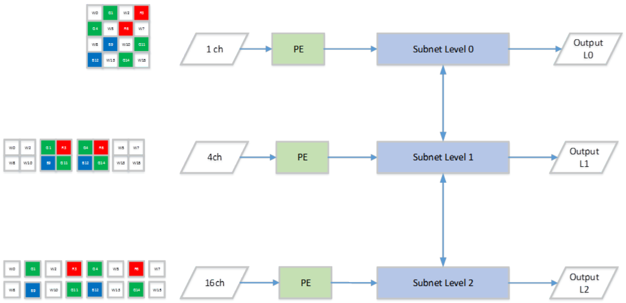

Figure 2. DePhaseNet

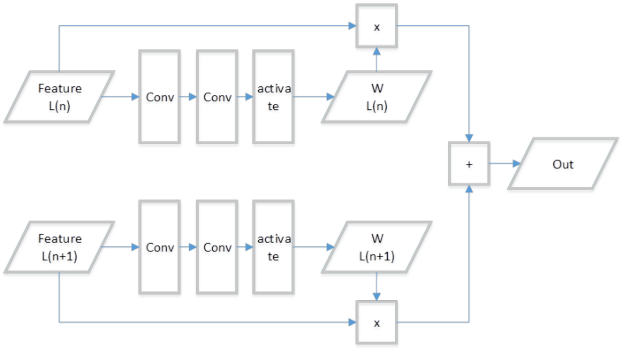

Figure 3. Distributed attention module

Weight distribution is close to zero mean distribution so that symmetric quantization that we use is effective in weight quantization.

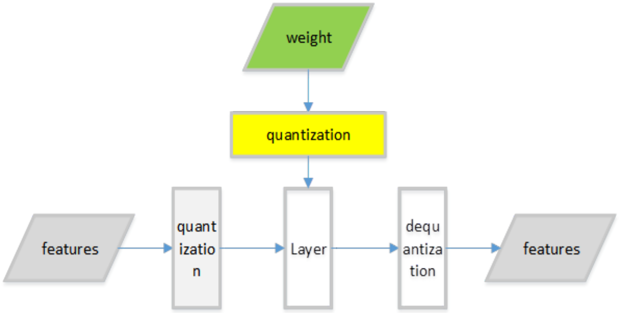

Figure 4. Quantization in deep network

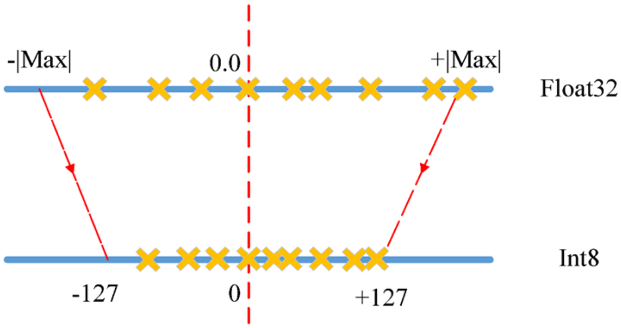

Figure 5. Symmetric quantization

Experimental results

We made experiments by preparing pairs of RGBW CFA pattern images and ground truth RGB images. The network was trained on MIT dataset and HDR+ Burst Photography Dataset [36]. We measured our algorithm on Kodak dataset [37] and real RGBW-K (kodak) image.

In Table. 1 and Fig. 6, objective image quality evaluation results on various bits are provided. 10 bit shows the image quality close to floating point. And 8 bit shows slight image quality degradation, 0.43 dB compared to floating point results.

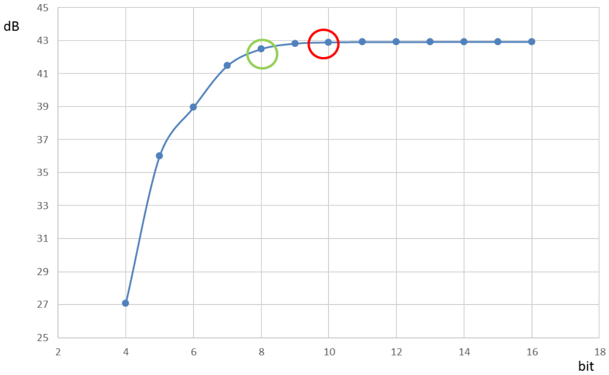

Figure 6. Quantized network output results on Kodak image, PSNR [dB]

Results for weight quantization in Kodak dataset, PSNR [dB]

Bit	PSNR [dB]
4	27.079
5	36.003
6	38.945
7	41.467
8	42.485
9	42.819
10	42.889
11	42.913
12	42.915
13	42.916
14	42.916
15	42.917
16	42.917
float	42.917

Subjective evaluation of experimental results show that we could see more quantization noises are shown in lower bit depth.

In kodak dataset, difference between 8 bit and 10 bit is slightly noticeable and 4 bit is worst in image quality as shown in Fig. 7 and Fig. 8.

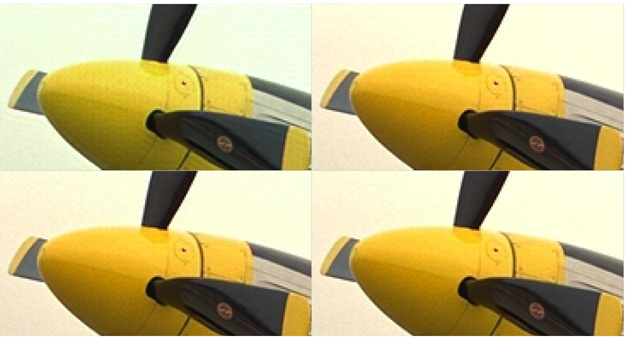

Figure 7. *Quantized network output results on Kodak image number 20: (a) - 4 bit; (b) - 6 bit; (c) - 8 bit; (d) - 10 bit.*

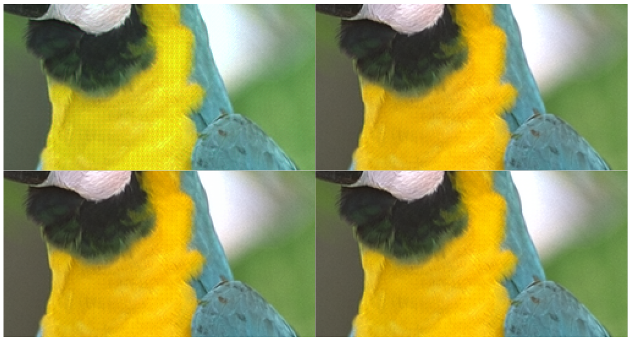

Figure 8. *Quantized network output results on Kodak image number 23: (a) - 4 bit; (b) - 6 bit; (c) - 8 bit; (d) - 10 bit.*

And we made tests on 10 bit real RGBW-K raw images, and we could see clear differences as shown in Fig. 9 and Fig. 10. From 8 bit to lower bits the quantized network results show various noticeable artifacts while we could see perfect image reconstruction when it comes to 10 bit. In Fig. 11, 8 bit shows artifacts on the letter ’o’ and ’n’. 6 bit and 4 bit show color change as well as artifacts.

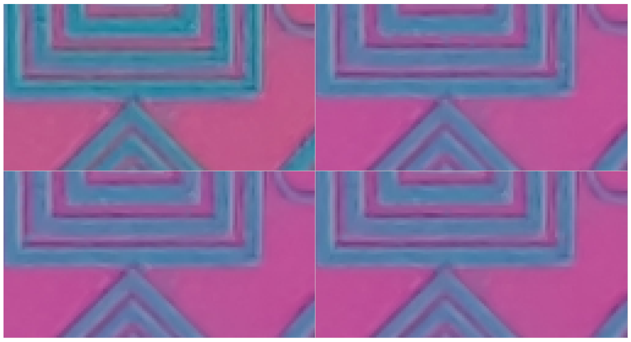

Figure 9. *Quantized network output results on real RGBW-K raw image 1: (a) - 4 bit; (b) - 6 bit; (c) - 8 bit; (d) - 10 bit.*

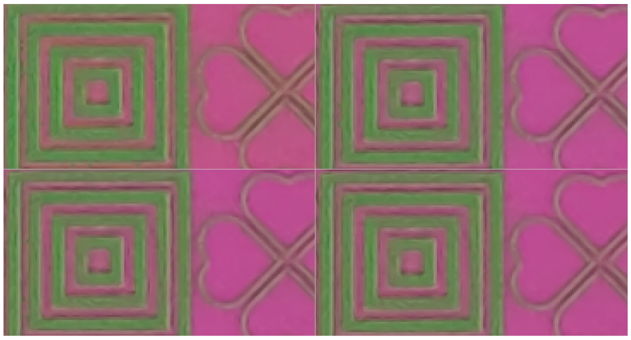

Figure 10. *Quantized network output results on real RGBW-K raw image 2: (a) - 4 bit; (b) - 6 bit; (c) - 8 bit; (d) - 10 bit.*

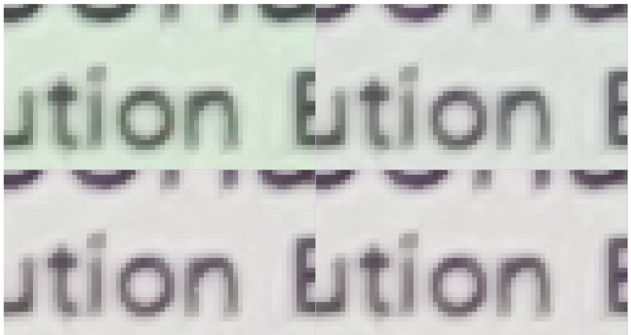

Figure 11. *Quantized network output results on real RGBW-K raw image 3: (a) - 4 bit; (b) - 6 bit; (c) - 8 bit; (d) - 10 bit.*

Conclusion

In this work, we studied quantization on deep learning network for RGBW-K demosaicing as a type of image restoration. Our research shows that 10 bit is most adequate for quantization of demosaicing both in objective quality and subjective quality aspect.

If one should deploy the network on usual AI hardware like NPU, DSP or GPU with conventional platform such as Tensorflow or Pytorch, only 8 bit and 16 bit are supported so that network should be quantized with 16 bit weights for best quality in the image restoration applications. In cases some artifacts and sligh image quality degradation are allowed or artifact reduction post-processing like denoising can be used, 8 bit may be used.

When it comes to the custom AI hardware or dedicated hardware, 10 bit is best choice for the weight and will make image quality close to floating point.

Recently some AI hardware platforms support float16, but it requires 1.5x area and power than fixed point hardware, so that integer quantized hardwares have still benefits.

References

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ”Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems 25 (2012): 1097-1105.

[2] Simonyan, Karen, and Andrew Zisserman. ”Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556 (2014).

[3] Denil, Misha, et al. ”Predicting parameters in deep learning.” arXiv preprint arXiv:1306.0543 (2013).

[4] Achterhold, Jan, et al. ”Variational network quantization.” International Conference on Learning Representations. 2018.

[5] LeCun, Yann, John S. Denker, and Sara A. Solla. ”Optimal brain damage.” Advances in neural information processing systems. 1990.

[6] Sze, Vivienne, et al. ”Efficient processing of deep neural networks: A tutorial and survey.” Proceedings of the IEEE 105.12 (2017): 2295-2329.

[7] Wang, Peisong, et al. ”Towards accurate post-training network quantization via bit-split and stitching.” International Conference on Machine Learning. PMLR, 2020.

[8] Gong, Yunchao, et al. ”Compressing deep convolutional networks using vector quantization.” arXiv preprint arXiv:1412.6115 (2014).

[9] Yang, Jiwei, et al. ”Quantization networks.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[10] Courbariaux, Matthieu, Yoshua Bengio, and Jean-Pierre David. ”Binaryconnect: Training deep neural networks with binary weights during propagations.” Advances in neural information processing systems. 2015.

[11] Rastegari, Mohammad, et al. ”Xnor-net: Imagenet classification using binary convolutional neural networks.” European conference on computer vision. Springer, Cham, 2016.

[12] Li, Fengfu, Bo Zhang, and Bin Liu. ”Ternary weight networks.” arXiv preprint arXiv:1605.04711 (2016).

[13] Zhu, Chenzhuo, et al. ”Trained ternary quantization.” arXiv preprint arXiv:1612.01064 (2016).

[14] Zhou, Shuchang, et al. ”Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients.” arXiv preprint arXiv:1606.06160 (2016).

[15] Li, Xin, Bahadir Gunturk, and Lei Zhang. ”Image demosaicing: A systematic survey.” Visual Communications and Image Processing 2008. Vol. 6822. International Society for Optics and Photonics, 2008.

[16] Menon, Daniele, and Giancarlo Calvagno. ”Color image demosaicking: An overview.” Signal Processing: Image Communication 26.8-9 (2011): 518-533.

[17] Cok, David R. ”Signal processing method and apparatus for producing interpolated chrominance values in a sampled color image signal.” U.S. Patent No. 4,642,678. 10 Feb. 1987.

[18] Adams Jr, James E. ”Interactions between color plane interpolation and other image processing functions in electronic photography.” Cameras and Systems for Electronic Photography and Scientific Imaging. Vol. 2416. International Society for Optics and Photonics, 1995.

[19] Adams Jr, James E. ”Interactions between color plane interpolation and other image processing functions in electronic photography.” Cameras and Systems for Electronic Photography and Scientific Imaging. Vol. 2416. International Society for Optics and Photonics, 1995.

[20] Dubois, Eric. ”Frequency-domain methods for demosaicking of Bayer-sampled color images.” IEEE Signal Processing Letters 12.12 (2005): 847-850.

[21] Hao, Pengwei, et al. ”A geometric method for optimal design of color filter arrays.” IEEE Transactions on Image Processing 20.3 (2010): 709-722.

[22] Mukherjee, Jayanta, R. Parthasarathi, and Sachin Goyal. ”Markov random field processing for color demosaicing.” Pattern Recognition Letters 22.3-4 (2001): 339-351.

[23] Keren, Daniel, and Margarita Osadchy. ”Restoring subsampled color images.” Machine Vision and applications 11.4 (1999): 197-202.

[24] Zhang, Chao, et al. ”Universal demosaicking of color filter arrays.” IEEE Transactions on Image Processing 25.11 (2016): 5173-5186.

[25] Gharbi, Michaël, et al. ”Deep joint demosaicking and denoising.” ACM Transactions on Graphics (ToG) 35.6 (2016): 1-12.

[26] Tan, Runjie, et al. ”Color image demosaicking via deep residual learning.” IEEE Int. Conf. Multimedia and Expo (ICME). Vol. 2. No. 4. 2017.

[27] Tan, Daniel Stanley, Wei-Yang Chen, and Kai-Lung Hua. ”DeepDemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks.” IEEE Transactions on Image Processing 27.5 (2018): 2408-2419.

[28] Syu, Nai-Sheng, Yu-Sheng Chen, and Yung-Yu Chuang. ”Learning deep convolutional networks for demosaicing.” arXiv preprint arXiv:1802.03769 (2018).

[29] Kim, Irina, et al. ”On recent results in demosaicing of Samsung 108MP CMOS sensor using deep learning.” 2021 IEEE Region 10 Symposium (TENSYMP). IEEE, 2021.

[30] Kim, Irina, et al. ”Under display camera quad bayer raw image restoration using deep learning.” Electronic Imaging 2021.7 (2021): 67-1.

[31] Banner, Ron, et al. ”Post-training 4-bit quantization of convolution networks for rapid-deployment.” arXiv preprint arXiv:1810.05723 (2018).

[32] Jacob, Benoit, et al. ”Quantization and training of neural networks for efficient integer-arithmetic-only inference.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[33] Chung, Kuo-Liang, Tzu-Hsien Chan, and Szu-Ni Chen. ”Effective three-stage demosaicking method for RGBW CFA images using the iterative error-compensation based approach.” Sensors 20.14 (2020): 3908.

[34] Kwan, Chiman, and Jude Larkin. ”Demosaicing of bayer and CFA 2.0 patterns for low lighting images.” Electronics 8.12 (2019): 1444.

[35] Kwan, Chiman, and Bryan Chou. ”Further improvement of debayering performance of RGBW color filter arrays using deep learning and pansharpening techniques.” Journal of Imaging 5.8 (2019): 68.

[36] Hasinoff, Samuel W., et al. ”Burst photography for high dynamic range and low-light imaging on mobile cameras.” ACM Transactions on Graphics (ToG) 35.6 (2016): 1-12.

[37] Loui, Alexander, et al. ”Kodak’s consumer video benchmark data set: concept definition and annotation.” Proceedings of the international workshop on Workshop on multimedia information retrieval. 2007.

Author Biography

Youngil Seo received his B.S in Electrical Engineering from Hanyang University and M.S. in Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2001 and 2003, respectively. From 2001 to 2009, he developed telemetics system in LG Electronics. Since 2009, he has been with Samsung Electronics where he developed Video codec, GPU, Sensor IP and so on. His main research interests include image processing systems and deep learning now.

Irina Kim received her B.S. and M.S. degrees with honors in applied mathematics, in 2002 and 2004, respectively, and the Ph.D. course in mathematics from the National Research University of Electronic Technology, Moscow, in 2005. Since 2001, she developed Nano microscopic and satellite image analysis and video surveillance algorithms, before joining Samsung Electronics in 2005, where she worked on Image Signal Processor (ISP) algorithms, face detection and vision engines. Her recent research is focused on deep learning for latest CMOS sensor for mobile devices.

Seongwook Song received his B.S. and M.S. degrees in electrical engineering from Seoul National University, in 1997 and 1999, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, in 2004. He has been with Samsung Electronics since 2003, to develop 2G, 3G and 4G chipsets. His main research interests include advanced signal processing for digital communications, multimedia and deep learning systems for digital cameras.