Supplementary Materials for AQD: Towards Accurate Quantized Object Detection

Peng Chen[2*] Jing Liu[1*] Bohan Zhuang[1†] Mingkui Tan[3] Chunhua Shen[1,2]
[1]Monash University [2]University of Adelaide [3]South China University of Technology

S1. More Results on ImageNet

Implementation details. Following HAQ [8], we quantize all the layers, in which the first and the last layers are quantized to 8-bit. Following [4, 2], we introduce weight normalization during training. We use SGD with nesterov [6] for optimization, with a momentum of 0.9. For all models on ImageNet, we first train the full-precision models and then use the pre-trained weights to initialize the quantized models. We then fine-tune for 150 epochs. The learning rate starts at 0.01 and decays with cosine annealing [5].

Main Results. We apply the proposed method to quantize MobileNetV1 [3] and MobileNetV2 [7] to 4-bit. We compare the performance of different methods in Table S1. From the results, our proposed method outperforms other methods by a large margin. For example, compared with HAQ, our proposed method achieve 2.7% and 3.5% higher Top-1 accuracy for 4-bit MobileNetV1 and MobileNetV2.

Table S1 – Performance comparisons on ImageNet.

Network	Method	Top-1 Acc. (%)	Top-5 Acc. (%)
MobileNetV1	Full-precision	70.9	89.8
	PACT [1]	62.4	84.2
	HAQ [8]	67.4	87.9
	Ours	70.1	89.3
MobileNetV2	Full-precision	71.9	90.3
	PACT [1]	61.4	83.7
	HAQ [8]	67.0	87.3
	Ours	70.5	89.5

References

[1] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 1

[2] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *Proc. Int. Conf. Learn. Repren.*, 2020. 1

[3] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[4] Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *Proc. Int. Conf. Learn. Repren.*, 2020. 1

[5] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *Proc. Int. Conf. Learn. Repren.*, 2017. 1

[6] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Proceedings of the USSR Academy of Sciences*, volume 269, pages 543–547, 1983. 1

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4510–4520, 2018. 1

[8] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *CVPR*, 2019. 1

[*] First two authors contributed equally.

[†] Corresponding author. E-mail: bohan.zhuang@monash.edu