Reducing the Side-Effects of Oscillations in Training of Quantized YOLO Networks – Supplementary Material –

Here, we provide additional experimental results and analysis. First, we provide a comparison between vanilla baseline methods and results using QC on the baselines on YOLO5 and YOLO7. Later, we also provide ablation studies of various setups of QC and show the stability of EMA for the decay parameter (α).

1. Comparisons with vanilla baselines and QC based baselines

Table 1. *Comparison between* LSQ *[1], Oscillation dampening [2], and our proposed method after performing our* QC *for quantization-aware training using mAP metric for object detection task on the COCO dataset.*

Method	Ours (QC)	#-bit	YOLO5-n	YOLO5-s	YOLO7-tiny
Full-Precision	-	32-bit	28.0	37.4	37.5
LSQ [1]	✗	4-bit	20.6	32.4	32.9
	✓		22.6	33.3	34.1
Osc. Dampening [2]	✗		21.5	32.9	33.5
	✓		23.1	33.4	34.3
Ours (EMA)	✗		22.1	33.1	34.6
	✓		**23.8**	**34.0**	**35.2**
LSQ [1]	✗	3-bit	15.2	27.2	28.4
	✓		17.1	29.4	30.2
Osc. Dampening [2]	✗		16.4	27.5	29.2
	✓		17.9	29.6	30.5
Ours (EMA)	✗		16.4	28.5	30.3
	✓		**18.2**	**30.2**	**31.0**

Further, we also perform experiments to evaluate the effectiveness of QC on the baseline QAT methods such as LSQ [1] and Oscillation dampening [2] on object detection task on COCO dataset and the results are reported in Table 1. Our QC approach to correct the error induced due to oscillating weights and scale factors cannot only improve the detection performance of quantized models of EMA but also the baseline methods. Despite that, our combined approach with both EMA and QC outperforms all the baselines with QC consistently at 4-bit as well as 3-bit quantization on YOLO5 and YOLO7 variants.

2. Ablation on different QC setups

QC can correct the error induced due to oscillations after the quantization by employing the per-channel scale and shift correction factors. These scale factors can also be chosen per-tensor. To evaluate the effectiveness of different components of QC such as scale and shift correction factors, we provide ablation studies in Table 2. We also provide results on both per-tensor and per-channel setup of QC. It is important to note that both scale and shift parameters are equally important in both per-tensor and per-channel QC setup and neither alone can effectively reduce oscillation-based error. Also, even the simple per-tensor setup of QC improves the EMA performance but as expected it cannot meet the performance achieved by the per-channel QC setting.

3. Effect of varying decay factor in EMA

To check the stability of EMA to varying decay factors, we trained different 4-bit YOLO5-n models using COCO datasets and the comparisons are provided in Table 3. As shown, our EMA approach is quite stable to different decay parameters. EMA

Table 2. *Ablation studies of different* QC *setups, where either per-tensor or per-channel correction is performed varying whether to use* QC *scale or shift on YOLO5-n trained at 4-bit on COCO dataset.*

QC Setup	QC Scale	QC Shift	mAP
Per-tensor	✓	✗	22.2
	✗	✓	22.3
	✓	✓	22.6
Per-channel	✓	✗	23.5
	✗	✓	23.6
	✓	✓	**23.8**

Table 3. *Different values of decay parameter* (α) *in* EMA *for YOLO5-n trained at 4-bit on COCO dataset. Note,* EMA *is quite stable with respect to different decay parameters.*

Decay parameter (α)	mAP
0.0	20.6
0.9	21.5
0.99	**22.1**
0.999	**22.1**
0.9999	**22.1**

takes into into account $\approx 1-(1-\alpha)$ iterations to compute the average weights or scale factors. Typically, oscillations are consistent over ≥ 100 iterations, and taking an average of scale factors or weights over ≥ 100 iterations works effectively in mitigating the oscillation side-effects in the final QAT model.

4. Oscillation in scale factors with or without EMA

To show the effect of EMA on the quantization scale factors during the training, we also provide the plots for scale factors during the last 4K iterations of training with or without EMA in Fig. 1 and Fig. 2 for scale factors of weights and activations respectively. It can be seen that EMA leads to a smoother transition of quantization scale factors for both weights and activations throughout the training and thus lead to stable training of quantized YOLO models.

References

[1] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 4

[2] Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. In *International Conference on Machine Learning*, pages 16318–16330. PMLR, 2022. 1

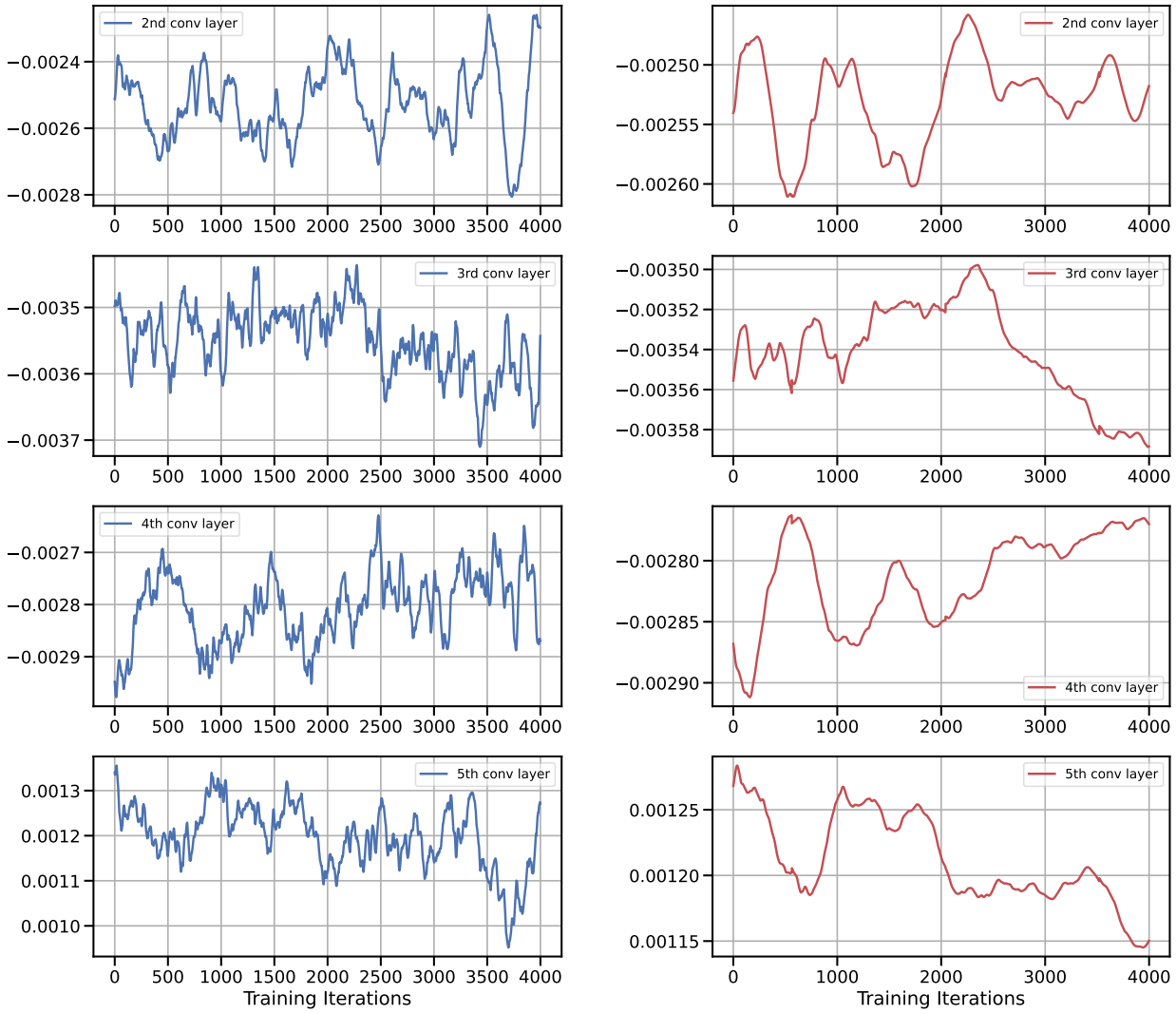

Figure 1. *Effect of* EMA *on oscillation issue in YOLO5-n variant trained on COCO dataset at 4-bit precision using* LSQ *[1]. (a) Scale factors for weight quantization in the vanilla model during the last 4K iterations of 2nd-5th conv layer, (b) Scale factors for weight quantization in* EMA *model during the last 4K iterations of 2nd-5th conv layer. Here, it can be observed that* EMA *makes the quantization scale factors stable for all the layers and gets rid of the oscillation issue.*

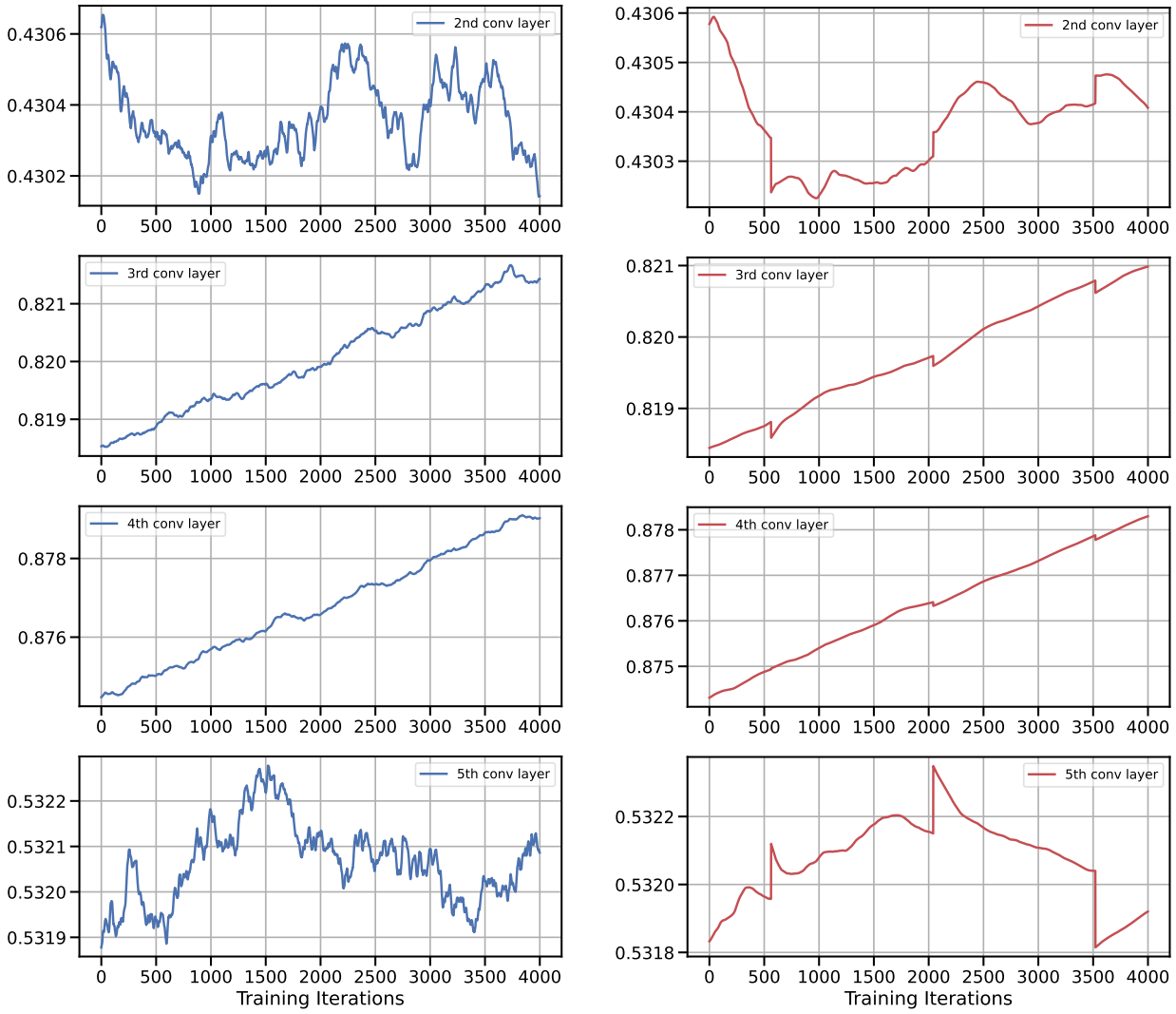

Figure 2. *Effect of* EMA *on oscillation issue in YOLO5-n variant trained on COCO dataset at 4-bit precision using* LSQ *[1]. (a) Scale factors for activation quantization in the vanilla model during the last 4K iterations of 2nd-5th conv layer, (b) Scale factors for activation quantization in* EMA *model during the last 4K iterations of 2nd-5th conv layer. Here, it can be observed that* EMA *makes the quantization scale factors stable for all the layers and gets rid of the oscillation issue.*