Supplementary Materials for
It's All In the Teacher: Zero-Shot Quantization Brought Closer to the Teacher

Kanghyun Choi[1], Hye Yoon Lee[1], Deokki Hong[1], Joonsang Yu[2],
Noseong Park[1], Youngsok Kim[1], and Jinho Lee[1]*
[1]College of Computing, Yonsei University [2]CLOVA ImageVision, CLOVA AI Lab, NAVER
[1]{kanghyun.choi,hylee817,dk.hong,noseong,youngsok,leejinho}@yonsei.ac.kr
[2]joonsang.yu@navercorp.com

A. Code

As a part of the supplementary materials, the code used to conduct experiments is attached as a separate zip archive. The zip archive contains the implementation of the AIT method on multiple zero-shot quantization backbones: GDFQ [4], ARC [6], and Qimera [1]. For reproducibility, the experiment environment setting and training scripts are included for all backbones. The code is under the terms of the GNU General Public License v3.0.

B. Lower Bit-width Experiments

Further experiments on GDFQ and ARC were conducted in lower-bit settings. The experiment results are shown in Tab. 1 Following the main paper, the ResNet family and MobileNet are denoted as 'RN' and 'MB', respectively. We tested 3w3a and 3w4a quantization settings for the ImageNet experiments and further down to 2w2a and 2w3a for Cifar-10/100, which we found to be the lowest bits GDFQ and AIT converge.

Dataset	Model	Bits	GDFQ		ARC	
			Baseline	AIT	Baseline	AIT
ImageNet	RN-18	3w3a	20.69	36.34	1.00	36.70
		3w4a	39.73	53.55	2.54	56.77
	RN-50	3w3a	0.21	1.31	0.20	3.98
		3w4a	26.85	37.50	1.37	49.34
	MB-V2	3w3a	5.50	13.83	0.20	30.35
		3w4a	26.87	37.77	0.22	47.41
Cifar-100	RN-20	2w2a	1.41	2.09	1.35	1.55
		2w3a	1.04	1.13	1.25	1.14
		3w3a	49.62	48.64	28.54	34.39
		3w4a	59.70	61.37	50.47	58.65
Cifar-10	RN-20	2w2a	16.48	15.57	16.18	13.47
		2w3a	37.64	40.98	20.87	20.42
		3w3a	80.70	80.49	52.99	51.78
		3w4a	90.02	90.20	82.10	82.98

Table 1. Low Bit-width Experiments Results

C. Experiments on Additional Network Models

We conducted a further evaluation of our method on various networks: InceptionV3 [3], SqueezeNext [2], and ShuffleNet [5]. The experimental results are shown in Table 2. Compared with the GDFQ baseline, our method still outperforms by a huge margin on all settings regardless of the quantization bitwidth. Furthermore, experimental results show that AIT is especially effective on smaller networks. This result again supports our observation in the main body that the limited capacity of a small network hinders the training phase from matching multiple loss terms simultaneously.

D. Comparison with Label Smoothing

Label smoothing is a regularization technique that replaces one-hot label y into a smooth label y' by

$$y' = (1-c)y + c/K, \tag{1}$$

where K is the number of classes and c is a label smoothing value. Label smoothing is known to help neural network training to avoid overfitting and increase generalization capability. Therefore, one might think that label smoothing can also help flatten the cross-entropy (CE) loss surface by its nature. To answer the question, we conducted comparative experiments with various label smoothing parameters. The experiments evaluate how the label smoothing affects the performance of GDFQ baseline and CE-only setting, which drops KL divergence from the training loss.

Dataset	Model (FP32 Acc.)	Bits	GDFQ	GDFQ +AIT
ImageNet	InceptionV3	4w4a	70.57	73.34 ($+2.77$)
	79.00	5w5a	77.25	77.67 ($+0.42$)
	SqueezeNext	4w4a	26.21	45.37 ($+19.16$)
	69.39	5w5a	56.07	62.76 ($+6.69$)
	ShuffleNet	4w4a	19.72	27.80 ($+8.08$)
	65.07	5w5a	45.92	48.97 ($+3.05$)

Table 2. Additional experiments on various network models.

---

*Corresponding author

ρ	Cifar-100	ImageNet	ρ	Cifar-100	ImageNet
	RN-20	RN-18		RN-20	RN-18
0.0005	65.41±0.20	64.48±0.28	0.00009	65.20±0.29	65.84±0.07
0.0004	65.55±0.15	65.23±0.10	0.00008	65.29±0.19	65.66±0.17
0.0003	65.44±0.34	65.41±0.53	0.00007	65.35±0.18	65.65±0.05
0.0002	65.21±0.27	65.85±0.07	0.00006	65.06±0.23	65.52±0.16
0.0001	65.04±0.13	65.51±0.09	0.00005	65.30±0.10	65.92±0.42

Table 3. Sensitivity Analysis on ρ.

Table 2 shows the experimental results. For CIFAR-10 and CIFAR-100, label smoothing did not improve performance in any settings over the baseline GDFQ, whether with KL divergence or not. Some improvements were observed from ImageNet dataset, but the improvements were smaller than that of AIT. This shows that even though label smoothing helps flatten the loss surface to some degree, its effect was not enough to reach that of AIT.

E. Further Analysis on ρ Sensitivity

We deepen the sensitivity analysis with finer levels of ρ values. The experiments are conducted five times per setting to demonstrate performance stability regarding ρ values. The results in Tab. 3 show that our method can achieve a stable accuracy level without hand-crafted hyperparameter tuning.

F. Gradient Cosine Similarity

Although the main body of the manuscript offers results for gradient cosine similarity measured on ResNet20 with CIFAR-10 dataset, we have done an extensive amount of experiments to study the distinct gradient directionality spotted in zero-shot quantization task. Here we share the results to further support our findings.

For CIFAR-10 and CIFAR-100 dataset, we used ResNet-20, ResNet-56, ResNeXt-29 32x4d, WRN28-10, and WRN40-8. On ImageNet, we evaluated on ResNet-18, ResNet-50, MobileNetV2, and InceptionV3. The experiment compares the directionality of loss functions in training these networks under two different settings: zero-shot quantization (ZQ) and knowledge distillation (KD). In the knowledge distillation setting, we used the same network for both the student and the teacher (self-distillation) for fair comparison against the Zero-shot quantization setting.

Fig. 1 shows the results for CIFAR-10, and Fig. 2 for CIFAR-100. Although the quantitative difference of cosine similarities and the details of its change throughout the training differs across different datasets and networks, one trend is consistent: KL divergence and cross-entropy disagrees with each other more under the zero-shot quantization setting. Such tendency is usually maintained throughout the training.

G. Hessian Trace

In this paper, Hessian matrix was used to measure the local curvature of the loss surface and compare the generalizability of the two distinct loss terms. Since Hessian matrix itself is enormous in size and computations involving its entirety is considered almost infeasible, analyzing the trace value of the matrix is often the most preferred way to study its characteristics. Adding to our results on Section 3.2 of the main body, we share further analysis on the loss curvature using Hessian trace.

We conducted further analysis on CIFAR-10 and CIFAR-100 datasets, on four different network models: ResNet-20, ResNet-56, WRN-28, and WRN-40. For all cases, our findings are the same. KL divergence has much smaller local curvature than the cross-entropy, where the gap is larger in zero-shot quantization settings.

References

[1] Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. In *Advances in Neural Information Processing Systems*, 2021. 1

[2] Amir Gholami, Kiseok Kwon, Bichen Wu, Zizheng Tai, Xiangyu Yue, Peter Jin, Sicheng Zhao, and Kurt Keutzer. Squeezenext: Hardware-aware neural network design. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2018. 1

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 1

[4] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *European Conference on Computer Vision*, 2020. 1

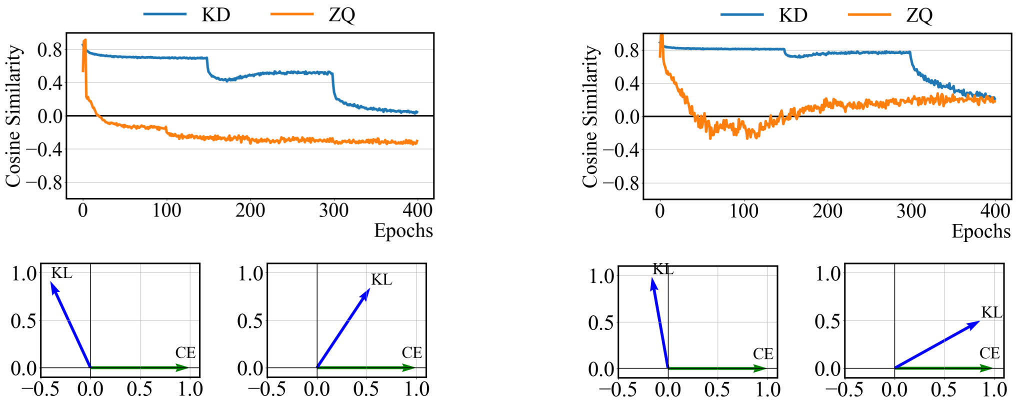
[5] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 1

[6] Baozhou Zhu, Peter Hofstee, Johan Peltenburg, Jinho Lee, and Zaid Alars. AutoReCon: Neural architecture search-based reconstruction for data-free compression. In *International Joint Conference on Artificial Intelligence*, 2021. 1

Dataset	Model	Method	c^*				AIT
			$0.00^\dagger$	0.10	0.30	0.50	
Cifar-10	ResNet-20	Baseline	90.25	89.67	88.85	88.52	91.23
		CE only	88.36	88.67	88.21	87.89	
Cifar-100	ResNet-20	Baseline	63.39	60.50	59.11	58.53	65.80
		CE only	56.76	60.10	59.13	57.81	
ImageNet	ResNet-18	Baseline	60.60	62.41	62.57	62.25	65.51
		CE only	60.33	62.48	62.27	62.18	

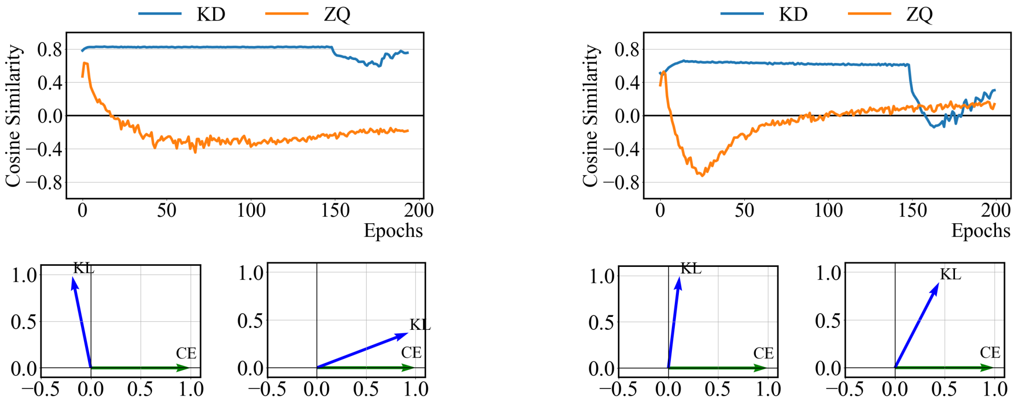
†No smoothing *Label smoothing parameter.

Table 4. Performance of GDFQ with label smoothing in 4w4a setting.



(a) CIFAR-10 ResNet-20

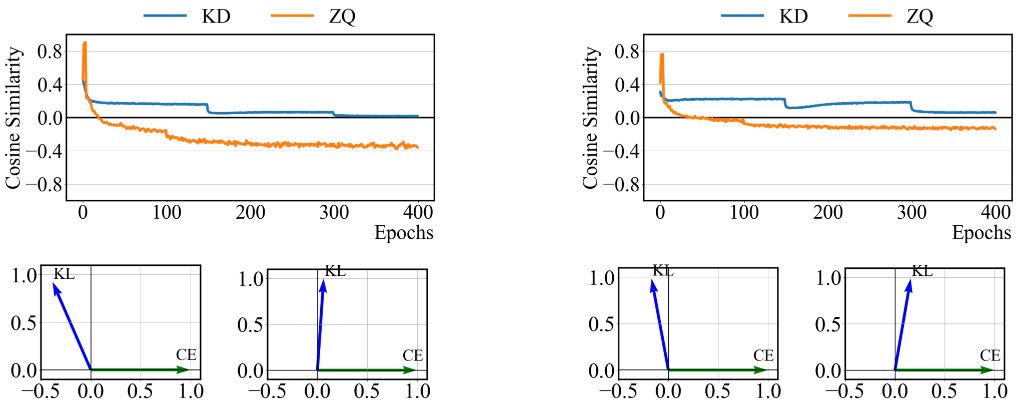
(b) CIFAR-10 ResNet-56



(c) CIFAR-10 WRN28-10

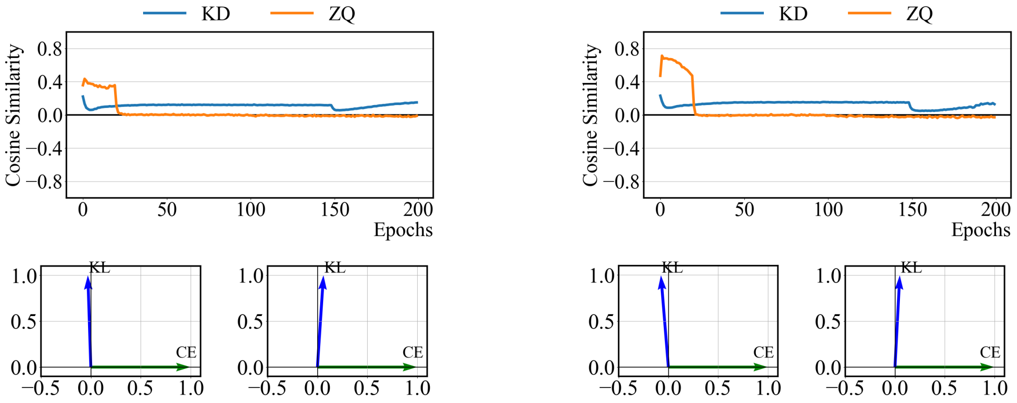
(d) CIFAR-10 WRN40-8

Figure 1. Gradient directionality of KL divergence and cross-entropy loss measured with CIFAR-10 dataset. In each setting, bottom left plots gradients under zero-shot quantization and bottom right plots gradients from knowledge distillation (self-distillation), captured from middle of the training.



(a) CIFAR-100 ResNet-20

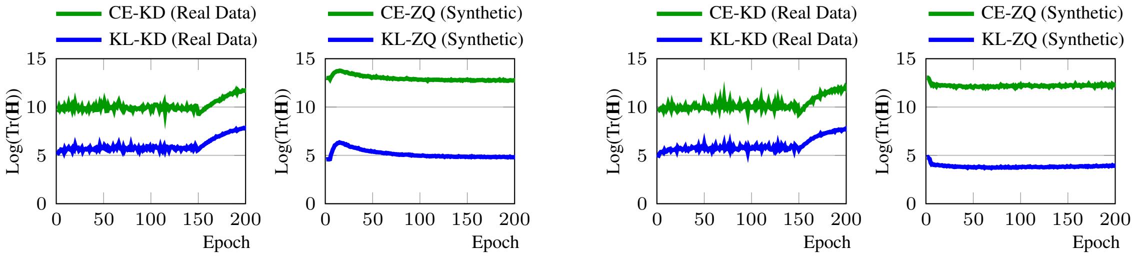
(b) CIFAR-100 ResNet-56



(c) CIFAR-100 WRN28-10

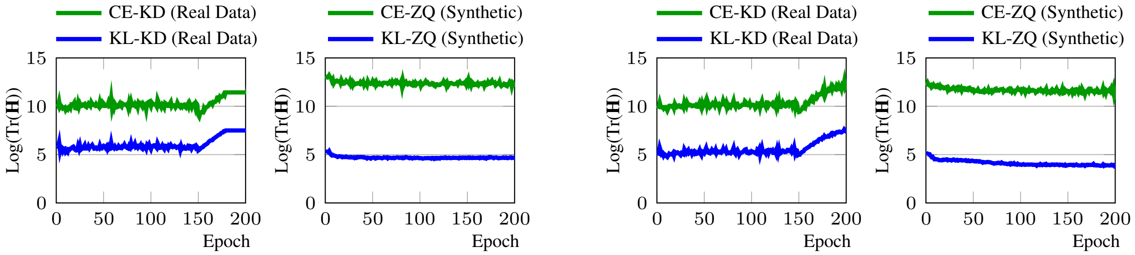
(d) CIFAR-100 WRN40-8

Figure 2. Gradient directionality of KL divergence and cross-entropy loss measured with CIFAR-100 dataset. In each setting, bottom left plots gradients under zero-shot quantization and bottom right plots gradients from knowledge distillation (self-distillation), captured from middle of the training.



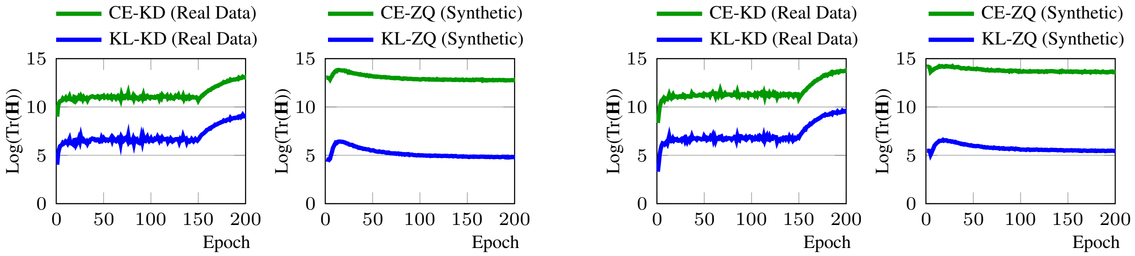
(a) CIFAR-10 ResNet-20

(b) CIFAR-10 ResNet-56



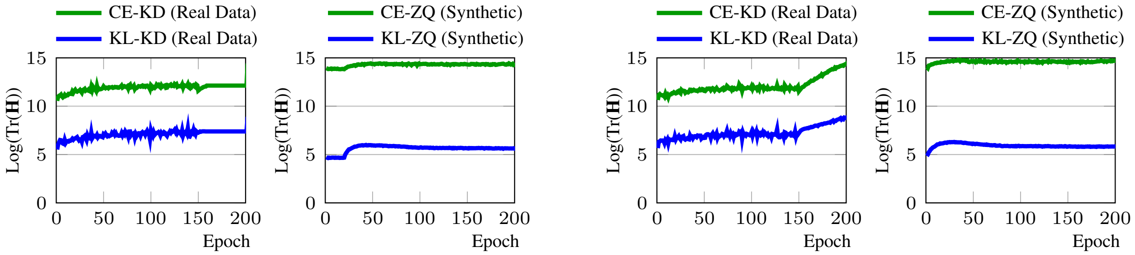
(c) CIFAR-10 WRN-28

(d) CIFAR-10 WRN-40



(e) CIFAR-100 ResNet-20

(f) CIFAR-100 ResNet-56



(g) CIFAR-100 WRN-28

(h) CIFAR-100 WRN-40

Figure 3. Hessian trace of KL divergence and cross-entropy, measured across diverse datasets and networks.