Hard Sample Matters a Lot in Zero-Shot Quantization
Supplemental Material

A. Sensitivity Analysis on Hyper-parameters

The hyper-parameters in our method including γ in Eq. 8, ϵ in Eq. 10 and λ in Eq. 12. In Figure 1, a sensitivity study of them on 3-bit ResNet-18 is performed. For γ, we choose several small values from 0 to 2 since the optimization in data generation process of ImageNet is not as easy as other small-scale datasets. We keep ϵ on a small magnitude so that the perturbation is not too large, and keep λ on a large magnitude so that the magnitude of two loss terms are consistent. The results show that the performance of HAST is somewhat sensitive to these hyper-parameters, but most of these results ($50.12\% \sim 51.15\%$) are comparable with that of the model fine-tuned on real data (51.95%). Note that the worse results in Figure 1 outperforms the quantized model obtained by the state-of-the-art ZSQ method (45.51%) in a large margin. We conduct similar experiments to find out the optimal value of these hyper-parameters on other datasets.

B. Sample Difficulty Promotion Details

Perturbation Direction Calculation. In the main paper, we calculate the perturbation δ by maximizing the sample difficulty, which is closely related to the loss. However, there are two loss terms, i.e., the Kullback-Leibler (KL) loss and the feature alignment (FA) loss in the fine-tuning process. Thus we conduct a further experiment to select the optimal loss for perturbation direction calculation. The experimental results are shown in Table 1. We observe that the choice of loss for calculating the perturbation direction has a certain impact on the performance. Though not optimal for all settings, we choose KL+FA to calculate perturbation direction since it shows the best in most settings.

loss weights. We apply sample difficulty promotion to the synthetic samples obtained by hard sample synthesis for more difficult samples. Then both of them are used to fine-tune the quantized model with the same loss weights. Further experiments on the loss weights of the original synthetic samples and the promotional samples are conducted. Experimental results are shown in table 2. The loss weight of the original synthetic samples is denoted as a, and that of the promotional samples is denoted as b. We perform 3-bit quantization on CIFAR-10 and ImageNet. For CIFAR-10, we achieve the best accuracy of 88.34% by setting both the weights to 1. When it comes to ImageNet, better performance than that reported in the main paper is obtained by increasing the weight of promotional samples.

Dataset	Model	Bit-width	KL	FA	KL+FA
Cifar-10	ResNet-20	W4A4	92.43	92.29	92.36
		W3A3	88.29	87.68	88.34
Cifar-100	ResNet-20	W4A4	66.69	66.50	66.68
		W3A3	55.61	55.13	55.67
ImageNet	ResNet-18	W4A4	66.90	66.69	66.91
		W3A3	51.06	50.87	51.15

Table 1. Performance of our HAST when calculating perturbation direction with diferent losses. We maximize the gradient of KL, FA and KL+FA respectively to calculate perturbation direction.

a, b	ResNet-20	ResNet-18	a, b	ResNet-20	ResNet-18
	Cifar-10	ImageNet		Cifar-10	ImageNet
1,0	86.17	47.94	0,1	88.19	48.55
3,1	85.92	50.52	1,4	86.69	52.14
2,1	87.53	50.97	1,3	86.94	53.12
1,1	88.34	51.15	1,2	87.73	52.69

Table 2. Ablation results of loss weights in W3A3 setting. The loss weights of original synthetic samples and promotional samples are denoted as a, b respectively.

C. Feature Alignment Analysis

Direct feature alignment vs. relaxed feature alignment. Direct feature alignment [3] is an easy and effective way to transfer feature representations by directly using mean square error to align the feature. However, we use attention vector [4] to relax the feature alignment constraint due to the limited capacity of quantized model. In this section, we provide the performance comparison of our HAST between using direct feature alignment (DFA) and using relaxed feature alignment (RFA). Table 3 shows the experimental results. The relaxed feature alignment obtains better performance in any settings over direct feature alignment. Significant improvements can be observed from 3-bit quantization. This

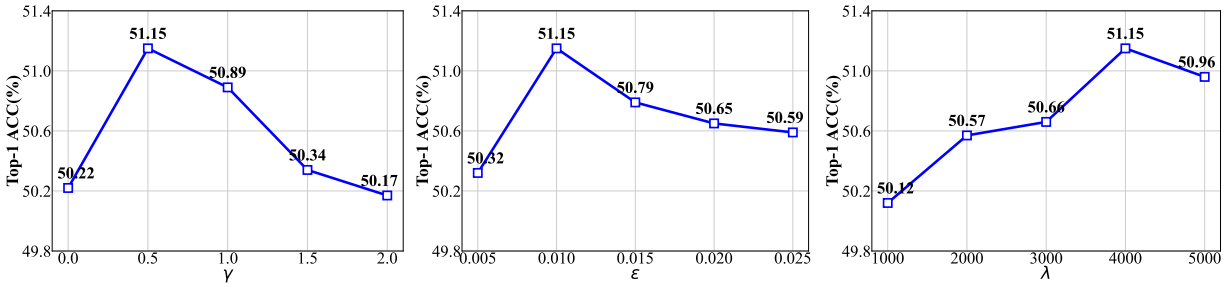
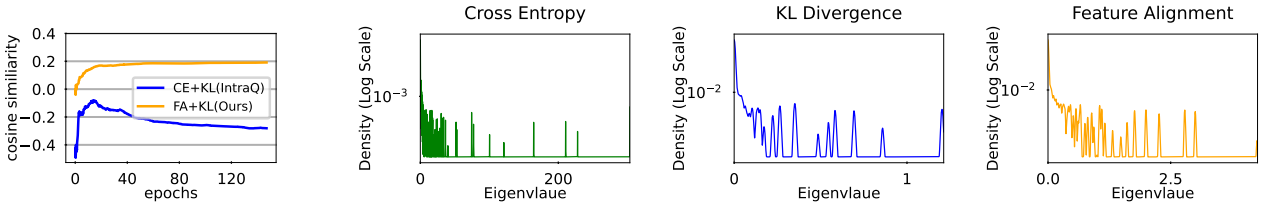


Figure 1. Sensitivity analysis on hyper-parameters. We report the top-1 accuracy of 3-bit ResNet-18 on ImageNet.



(a) Gradient cosine similarity.

(b) Distribution of eigenvalues.

Figure 2. Further experiments on feature alignment. (a)Gradient cosine similarity of two terms in loss function. (b)Distribution of the eigenvalues for different loss.

shows that it is harmful for low-precision quantized model to learn the feature representations of full-precision model directly.

Dataset	Model	Bit-width	HAST(DFA)	HAST(RFA)
Cifar-10	ResNet-20	W4A4	91.99	92.36
		W3A3	83.92	88.34
Cifar-100	ResNet-20	W4A4	66.53	66.68
		W3A3	51.50	55.67
ImageNet	ResNet-18	W4A4	66.49	66.91
		W3A3	45.52	51.15

Table 3. Performance of our HAST with direct feature alignment and relaxed feature alignment.

Cooperation with KL. Gradient cosine similarity was used in [1] to measure the cooperation ability of multiple loss terms. The authors found that the cross-entropy (CE) loss does not work well with the Kullback-Leibler (KL) loss in network fine-tuning process. We apply this metric in our work. Specifically, we fine-tune the 3-bit ResNet-20 using baseline (CE+KL) [5] and our HAST (FA+KL) respectively and measure the cosine similarity of the gradient of two distinct loss terms. As shown in Figure 2a, the cosine distance between CE and KL takes negative values throughout the fine-tuning, while that of FA+KL is positive. This implies that the combinations of FA and KL cooperate well, and using them together could enhance each other, which is opposite to the combinations of CE and KL.

Generalizability. Hessian matrix was used in [1] 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 eigenvalues of the matrix is often the most preferred way to study its characteristics. Figure 2b plots the distribution of the eigenvalues of the Hessian matrix, approximated by PyHessian [2]. We separate Hessian calculation for each loss of CE, KL and FA. A huge difference in the local curvature of the loss terms can be observed. While CE has longer tail for high eigenvalues, KL and FA has more concentration to lower eigenvalues, which means the local curvature of loss surface of KL and FA is smaller than that of CE, leading to better genrealizability according to the finding that smaller local curvature improves generalization [1].

D. Results with smaller number of samples

Table 4 shows the ablation on amount of the synthetic samples. The performance drops as the number of samples decreases. However, HAST with only 256 samples still performs better than previous methods, such as IntraQ with 45.51% using 5120 samples.

Amount	IntraQ(5120)	256	1280	2560	5120
ACC(%)	45.51	49.17	49.95	50.23	51.15

Table 4. Results with smaller number of samples.

References

[1] Kanghyun Choi, Hyeyoon Lee, Deokki Hong, Joonsang Yu, Noseong Park, Youngsok Kim, and Jinho Lee. It's all in the teacher: Zero-shot quantization brought closer to the teacher. In *CVPR*, 2022. 2

[2] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *IEEE International Conference on Big Data*, 2020. 2

[3] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 1

[4] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 1

[5] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *CVPR*, 2022. 2