Difficulty, Diversity, and Plausibility: Dynamic Data-Free Quantization

Cheeun Hong[1*] Sungyong Baik[3*] Junghun Oh[1] Kyoung Mu Lee[1,2]
[1] Dept. of ECE & ASRI, [2] IPAI, Seoul National University, Seoul, Korea
[3] Dept. of Data Science, Hanyang University, Seoul, Korea
cheeun914@snu.ac.kr, dsybaik@hanyang.ac.kr, dh6dh@snu.ac.kr, kyoungmu@snu.ac.kr

A. Analysis on bit-width candidates

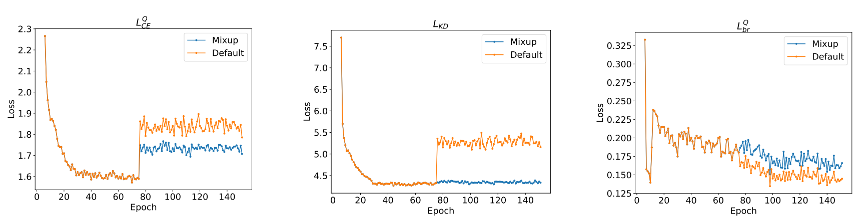
We investigate the impact of using different combinations of candidate bit-widths, as shown in Table S1. The results demonstrate that different combinations of candidates can provide flexibility in trade-offs between computational cost and accuracy. Compared to using three candidates ($M = 3$) $\{3, 4, 5\}$, adding candidates to ($M = 5$) for the bit selector results in an accuracy loss, but the bit-FLOPs are largely reduced. However, reducing the number of candidates ($M = 2$) leads to an accuracy drop with a small bit-flops reduction, which indicates that two candidates are insufficient to benefit from the dynamic framework.

Table S1. **Ablation study of DDPQ** of ∼4 MP on ImageNet with ResNet-18.

Bit-width candidates	Bit-FLOPs (%)	Top-1 (%)
{3, 4, 5}	2.50	67.47
{3, 5}	2.48	66.85
{2, 3, 4, 5, 6}	2.41	65.59

B. Analysis on mixup

We additionally analyze the impact of the mixup technique in Figure S1. The figure shows the curves of the loss terms during training. After 50% of epochs, when the mixup is applied, we compare the loss using the mixed-up images (blue line) and the loss value of the images before the mixup (orange line). The results show that mixup reduces the loss terms $\mathcal{L}_{ce}^{Q}$ (Figure S1a) and $\mathcal{L}_{kd}$ (Figure S1b). Notably, as shown in Figure S1c, bit regularization loss increases after mixup is applied and remains large for mixed-up images. The reason is that, since mixup is applied on easy samples to further promote the difficulty diversity the network is exposed to, the overall difficulty of the synthetic training set is enhanced. Thus, it is natural to leverage higher bit-widths for difficult samples, and therefore, the bit regularization loss term increases.



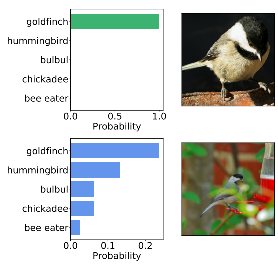
(a) $\mathcal{L}_{ce}^{Q}$ (b) $\mathcal{L}_{kd}$ (c) $\mathcal{L}_{br}$

Figure S1. **Learning curves of losses using mixup** for ImageNet using ResNet-18.

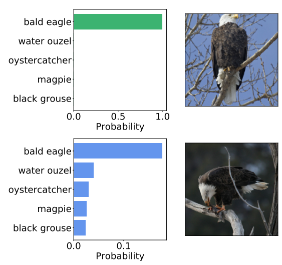
*equal contribution

C. Additional visualization

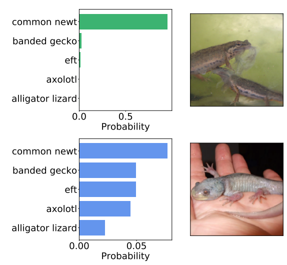
We provide additional visualizations to validate the generalizability of the observation of the main manuscript; difficult images are confusing between similar classes and the class of the classification layer can function as a prototype of that class. According to Figure S2 which displays more examples of difficult and easy images in the original dataset (ImageNet), difficult images tend to be confusing among semantically similar classes. We visualize the two images with the highest and lowest prediction entropy of each class. Moreover, we additionally validate the correlation between the similarity of class weights and the actual co-occurrences of classes in prediction output, as in Figure S3. The results demonstrate that we can estimate the semantic similarity between classes with weights of the classification layer without accessing the original data.



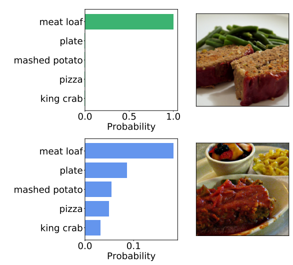
(a) Goldfinch



(b) Bald eagle

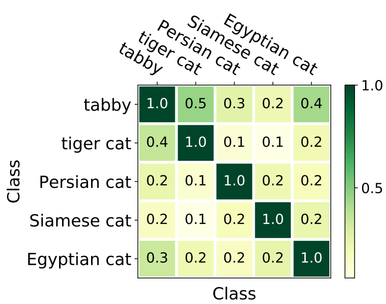


(c) Common newt

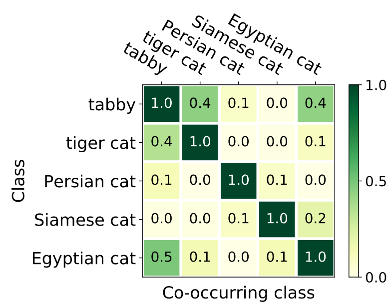


(d) Meat loaf

Figure S2. **More examples of easy and difficult images in ImageNet.**



(a) Classification weight similarity



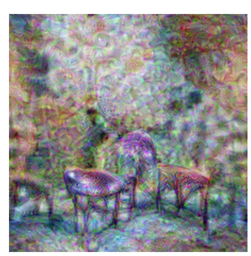
(b) Class co-occurrences in prediction

Figure S3. **Correlation between class weight similarity and class co-occurrences in prediction** on five cat classes of ImageNet.

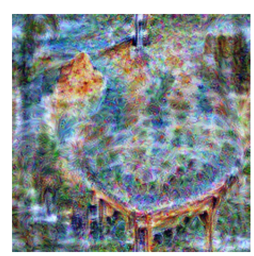
Furthermore, we investigate the impact of the hyper-parameter K which is the number of similar classes, on the generated images. In Fig. S4, the synthesized image of the same top-1 class becomes more complex for a larger K. This indicates that a large K can produce an overly confusing image due to the assigned irrelevant classes, such as *window screen* class, to generate a confusing *dining table* image. K=2 is sufficient to generate a confusing image for the classification network.

$K = 1$
(**dining table**)

$K = 2$
(**dining table**, patio)



$K = 3$
(**dining table**, patio, rocking chair)



$K = 10$
(**dining table**, patio, rocking chair, barber chair, china cabinet, window -screen, carriage, closet, window -shade, entertainment center)

Figure S4. **Generated samples of the same top-1 class (dining table) using different Ks.** Images are synthesized from ImageNet-pretrained ResNet-18. Classes in brackets denote the assigned top-K classes.