Non-Uniform Step Size Quantization for Accurate Post-Training Quantization: Supplementary Materials

Sangyun Oh[1], Hyeonuk Sim[2], Jounghyun Kim[3], and Jongeun Lee[1,3†]

[1] Department of Electrical Engineering, UNIST, Ulsan, Korea
[2] Department of Computer Science and Engineering, UNIST, Ulsan, Korea
[3] Artificial Intelligence Graduate School, UNIST, Ulsan, Korea
{syoh, detective, maxedset, jlee}@unist.ac.kr

1 Exploring Universal Set Design

We described the design of universal set in Section 3.4 of the paper and chose Option 5 from Table 1. Since each option represents a different set of quantization points, the choice of an option directly affects the composition of universal set (S_U), which has a profound impact on quantization error and model accuracy as well as hardware complexity. Figure 1 visualizes the universal sets of the five options we considered.

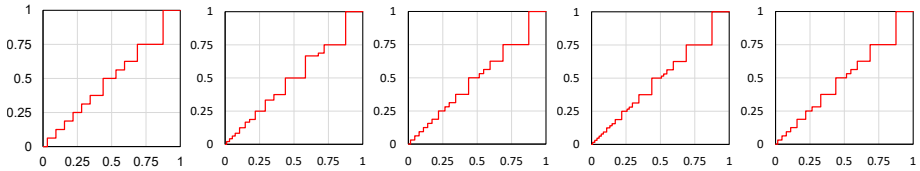
Table 1: Exploring universal set design, $S_U = \{a + b \,|\, a \in A, b \in B\}$.

Option	A	B
1	$\{2^{-1}, 2^{-2}, 2^{-3}, 0\}$	$\{2^{-1}, 2^{-2}, 2^{-4}, 0\}$
2	$\{1, 2^{-2}, 2^{-4}, 0\}$	$\{2^{-1}, 2^{-3}, 2^{-5}, 0\}$
3	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 0\}$	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 0\}$
4	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 0\}$	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 0\}$
5 (chosen)	$\{1, 2^{-1}, 2^{-3}, 0\}$	$\{1, 2^{-2}, 2^{-4}, 0\}$

We use the ImageNet dataset and ResNet-18 model to evaluate the five options. The result is summarized in Table 2. An important attribute of a universal set is the number of combinations that it generates, since our algorithm searches all combinations $\binom{|S_U|}{N}$ where N represents the representational precision. For example, Option 4 has the largest number of combinations ($= 490,314$) and can expect a smallest weight quantization error on average; however, it also has the longest QPS search time (almost 2 days) and the highest hardware cost.

In addition, after determining the best QPS for each layer (i.e., QPS search), we optimize the scale factor(s) using a calibration set, so that weights can be partially re-mapped to better quantization points within each QPS according to activation. Because of this additional step, smaller weight quantization error does

† Corresponding author.



(a) Option 1 (b) Option 2 (c) Option 3 (d) Option 4 (e) Option 5

Fig. 1: Distribution of quantization points (red points) of universal set for each option.

not necessarily mean smaller accuracy drop. Our result in Table 2 does suggest that Option 4 indeed has the highest accuracy when only weights are quantized. However, in terms of the final quantization accuracy using a calibration set, Option 3 shows the highest accuracy.

To estimate the *relative* hardware cost of different options we make the following assumptions. First, we consider only the shifters part, which is the multiplier part minus the adder. Though different options will lead to different adder and accumulator sizes, for our comparison between the five options, a more complicated multiplier also generates wider multiplication results, requiring wider adder and accumulator. Because of this positive correlation, considering the shifters' complexity is sufficient. Second, we assume that the shifters part is implemented with two MUXes and constant-amount shifters, as illustrated in Figure 2. Constant-amount shifter is just bit selection (wires only), consuming no logic gates.

Our hardware cost model is as follows: Cost of k-bit M-to-1 MUX $= k(M-1)$, which models the number of 1-bit 2-to-1 MUXes needed. In practice, various logic optimizations will reduce the complexity of design differently (due to guaranteed zero values in certain positions), which is ignored in our model. The value of M is evident from the option definition or from Figure 2. The value of k is the width of x plus the maximum shift amount, which may be different between Shifter 1 and Shifter 2.

We have finally selected Option 5, which has a reasonable QPS search time of about 0.5 hour for ResNet-18 and also shows sufficiently high accuracy comparable to Options 2 and 3, and even higher than Option 4. Also importantly, Option 5 has the lowest hardware cost together with Option 1.

2 Visualization

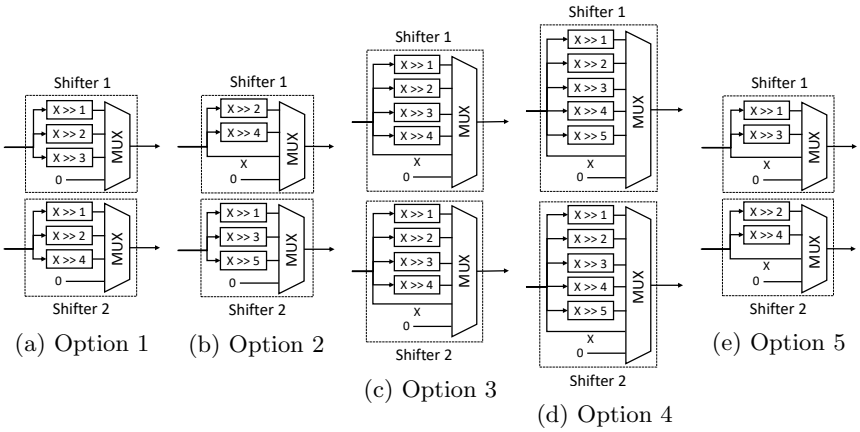
We provide a visualization example of the results we presented in the paper.

2.1 Object Detection

For object detection, we have obtained results using the Pascal VOC 2007 test set and the baseline model is SSD-lite (mAP 70.84%) [2] with MobileNetV2 [3]

Table 2: Exploration result of universal set design (*: chosen).

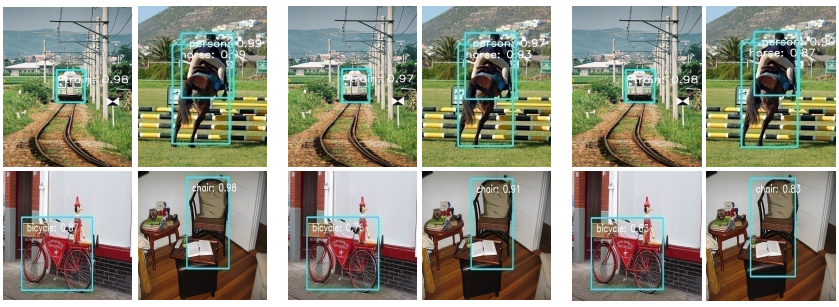
Option	1	2	3	4	**5***
$\|S_U\|$	12	16	17	23	15
QPS Size	495	12,870	24,310	490,314	6,435
QPS search time	**0.03h**	1h	2h	40h	0.5h
Relative HW cost (4-bit input)	**45**	51	80	108	**45**
Top-1 (wgt. only)	67.14	68.63	68.57	**68.69**	68.50
Top-1 (final opt.)	68.79	69.49	**69.51**	69.39	69.46



(a) Option 1 (b) Option 2 (c) Option 3 (d) Option 4 (e) Option 5

Fig. 2: Shifters design (multiplier minus an adder) for each option.

backbone. Figure 3 shows the visualized results. In both 4-bit and 3-bit cases, the boundary boxes are drawn very similarly to those of the baseline. However, as we have suggested in the paper, when weight precision is reduced to 3-bit, mIoU drops by about 3.8% point, but its detection performance is still very acceptable.

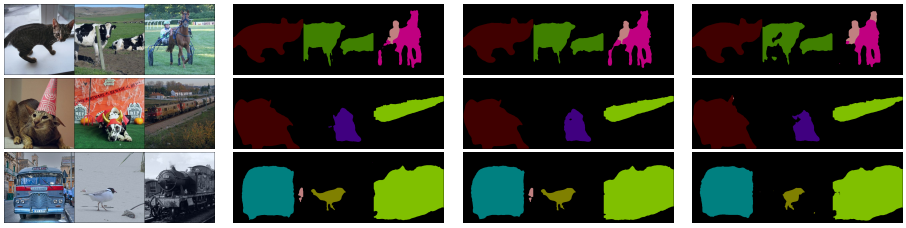


(a) FP32 (68.59%) (b) W4A8 (68.21%) (c) W3A8 (64.72%)

Fig. 3: Object detection results (mAP).

2.2 Semantic Segmentation

For semantic segmentation, we have obtained results using the Pascal VOC 2012 test set and the baseline model is DeepLabV3+ (mIoU 70.81%) [1] with MobileNetV2 [3] backbone. Figure 4 shows the visualized results. Overall our results show a similar trend as that of the object detection results. Our 4-bit results have similar quality as the baseline. At 3-bit we observe noise in the case of ambiguous images such as the bottom center case, but the overall segmentation performance is still quite acceptable.



(a) Orig. image (b) FP32 (70.81%) (c) W4A8 (68.96%) (d) W3A8 (61.94%)

Fig. 4: Semantic segmentation results (mIoU).

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. Eur. Conf. Comput. Vis. (ECCV) (2018)
2. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: Ssd: Single shot multibox detector. In: Eur. Conf. Comput. Vis. (ECCV) (2016)
3. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)