Supplementary Materials for: MSQ: Memory-Efficient Bit Sparsification Quantization

Seokho Han[1*] Seoyeon Yoon[1*] Jinhee Kim[1] Dongwei Wang[2]
Kang Eun Jeon[1✉] Huanrui Yang[2✉] Jong Hwan Ko[1✉]
[1]Department of Electrical and Computer Engineering, Sungkyunkwan University, Korea
[2]Department of Electrical and Computer Engineering, University of Arizona, USA
{beppa2396, syy000405, a2jinhee, kejeon, jhko}@skku.edu
{dongweiw, huanruiyang}@arizona.edu

1. Changes in Layer Precision During Training

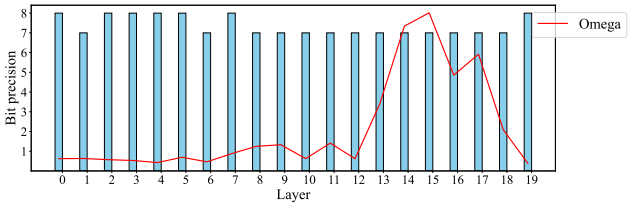
Fig. 1 illustrates how Omega values and bit precision change across layers during the training process of ResNet-20. Our Hessian Aware Aggressive Pruning method dynamically assigns prune bits as either 1-bit or 2-bit based on Omega values, enabling efficient bit reduction while maintaining model performance.

In the first pruning step Fig. 1a, Hessian Aware Aggressive Pruning is not yet applied, so all layers retain a prune bit of 1. After pruning occurs, the Hessian is calculated, and prune bits are dynamically reassigned as either 1-bit or 2-bit depending on Omega values.

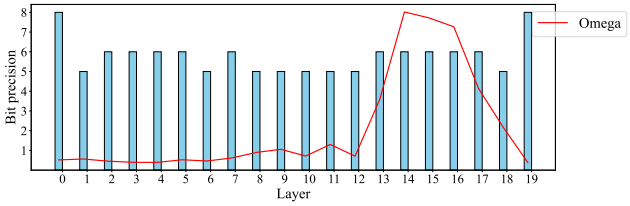
As training progresses to the second pruning step (Fig. 1b), pruning occurs across most layers. However, as observed in Fig. 1a and Fig. 1b, layers that were assigned a prune bit of 2 in the first step undergo a rapid reduction in bit precision. For example, Layer index 11 is reduced from 7-bit to 5-bit, whereas Layer index 13 decreases from 7-bit to 6-bit.

This iterative pruning step continues, progressively refining bit precision across layers. Ultimately, the model achieves a compression rate of 17.11×, as depicted in Fig. 1e.

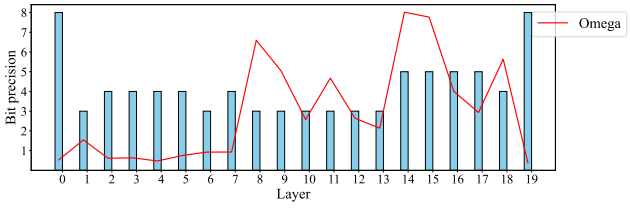
*: Equal contributions
✉: Corresponding authors



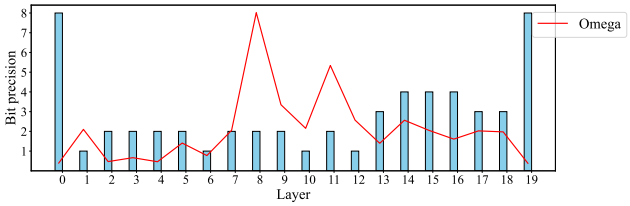
(a) First Pruning Step (Compression Rate: 4.54x)



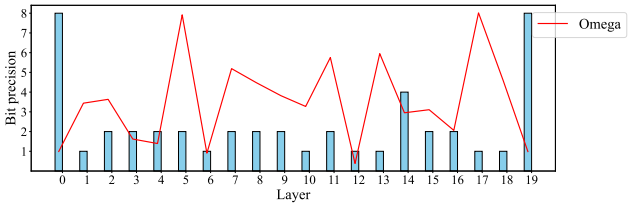
(b) Second Pruning Step (Compression Rate: 5.63x)



(c) Third Pruning Step (Compression Rate: 7.20x)

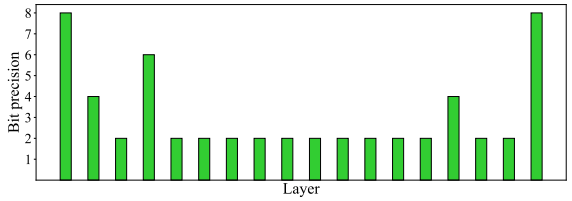


(d) Fourth Pruning Step (Compression Rate: 10.30x)

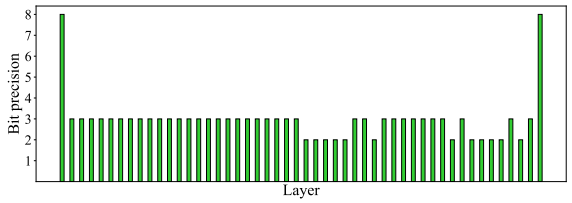


(e) Last Pruning Step (Compression Rate: 17.11x)

Figure 1. Changes in Layer Precision and Omega During Training on ResNet-20

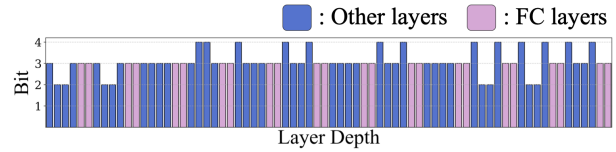


(a) ResNet-18

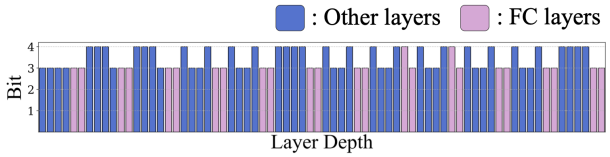


(b) ResNet-50

Figure 2. Final bit schemes of ResNet-18 and ResNet-50 after 100 epochs of training with MSQ.



(a) DeiT-T



(b) DeiT-S

Figure 3. Final bit schemes of DeiT-T and DeiT-S after fine-tuning with MSQ. The FC layers are colored in pink.

2. Final Bit Schemes

The final bit schemes for ResNet-18, ResNet-50, and ViT models, which were not included in the main text, are provided in this section. Fig. 2 presents the bit schemes for ResNet-18 and ResNet-50. Specifically, Fig. 2a achieves a compression ratio of 11.84×, whereas Fig. 2b achieves a compression ratio of 10.89×. Fig. 3 presents the bit schemes for DeiT-T and DeiT-S. Fig. 3a achieves a compression ratio of 10.54×, whereas Fig. 3b achieves a compression ratio of 9.58×. Notably, the fully connected (FC) layers, which are observed to be pruned more aggressively.

3. Additional Experiments on ViT

To further validate the scalability of our method on larger transformer-based models, we conduct supplementary experiments on ViT-Base-Patch16-224. While the main paper presents results on compact vision transformers such as DeiT-T, DeiT-S, and Swin-T, as shown in evaluate on a substantially larger Table 1 architecture with 86.6M parameters.

Table 1. Evaluation on ViT-Base-Patch16-224 using CIFAR-100.

Method	W-Bits	Comp(×)	Acc(%)
FP	32	1.00	92.06
DoReFa	4	8.00	90.20
MSQ	MP	**9.14**	**91.45**

Hyperparameters	
Epochs	50
λ	5e−5
I	5
α	0.3

4. Hyperparameter Details

The main hyperparameters of MSQ include λ, which controls the L1 regularization strength, α, the pruning threshold that determines pruning decisions based on each layer's LSB non-zero rate, and I, the pruning interval. The pruning interval I is crucial for guiding LSB sparsification and facilitating accuracy recovery after pruning. Our experimental settings can be found in Table 2.

Reducing λ decreases the regularization strength on LSBs, leading to less sparsity. Conversely, increasing λ strengthens the regularization effect, deriving higher sparsity as depicted on Fig. 4. However, setting λ too high may cause excessive LSB regularization, potentially degrading accuracy. Thus, it is essential to carefully tune λ and the pruning threshold α to balance sparsity and accuracy effectively.

Table 2. Hyperparameter settings for pruning on our experiments. λ represents the L1 regularization strength, α determines the pruning decision based on each layer's non-zero rate, and I denotes the pruning interval.

Network	λ	α	I
ResNet-20	5e-5	0.3	20
ResNet-18	5e-5	0.3	10
ResNet-50	5e-5	0.3	10
DeiT-T	8e-6	0.35	5
DeiT-S	5e-6	0.35	8
Swin-T	5e-6	0.35	8
MobileNetV3-Large	5e-5	0.3	5

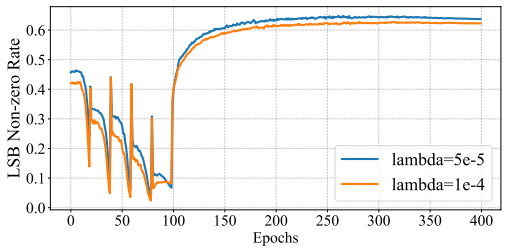


Figure 4. Comparison of $\lambda = 5e{-}5$ and $\lambda = 1e{-}4$. The LSB non-zero rate is relatively smaller when $\lambda = 1e{-}4$.