# Searching for Robust Binary Neural Networks via Bimodal Parameter Perturbation

Daehyun Ahn[*]
SqueezeBits Inc.
daehyun.ahn@squeezebits.com

Hyungjun Kim[*]
SqueezeBits Inc.
hyungjun.kim@squeezebits.com

Taesu Kim[*]
SqueezeBits Inc.
taesu.kim@squeezebits.com

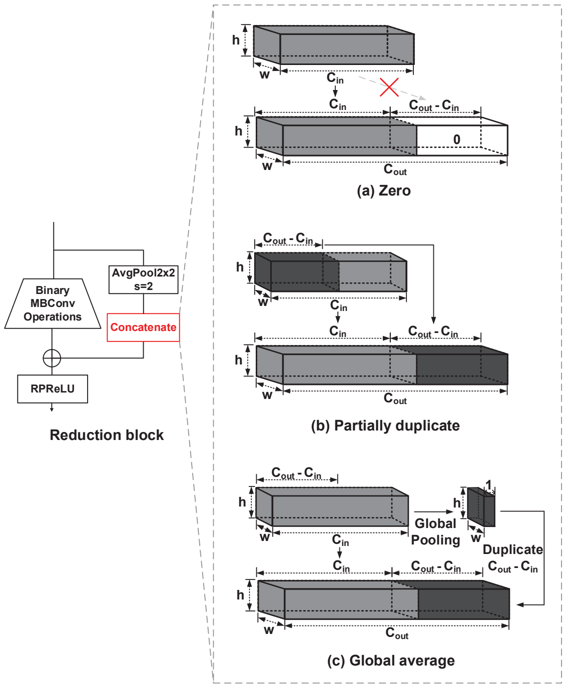
Eunhyeok Park
Pohang University of Science and Technology
eh.park@postech.ac.kr

Jae-Joon Kim
Seoul National University
kimjaejoon@snu.ac.kr

In the supplementary material, the contents which are not contained in the manuscript are described including the actual structures of searched BNN models and additional experimental results. First, Sec. 1 shows the experimental results on various shortcut path candidates for the reduction binary MBConv block. We analyze the search cost of our framework in Sec. 2. Structures of searched cells on BNAS and that of searched models based on binary MBConv block under various perturbation conditions are illustrated in Sec. 3 and Sec. 4. Lastly, we compare the performance of our searched BNN models on CIFAR dataset with that of previous works in Sec. 5.

## 1. Experiments on shortcut path in the reduction binary MBConv block

As we stated in Section 5.1 of the manuscript, the BNN models based on modified binary MBConv block failed to converge without any shortcut path in the reduction block. We proposed and tested three candidates of shortcut path without floating-point matrix-matrix multiplication illustrated in Fig. 1. Given the number of input channels $C_{in}$, the number of output channels $C_{out}$ in a reduction binary MBConv block and the output of AvgPool $X_{AP}$, the candidates are classified by the method to compute a matrix with $C_{out} - C_{in}$ channels which will be concatenated to $X_{AP}$ for the addition to the output of second $1 \times 1$ convolution. 'Zero' shortcut type (Fig. 1a) concatenates a zero matrix $O_{C_{in}-C_{out}}$ to $X_{AP}$, and partially duplicated matrix $X_{AP}[: C_{out} - C_{in}, :, :]$ is concatenated to $X_{AP}$ in 'Partially duplicate' type (Fig. 1b). In Fig. 1c ('Global average'), we compute the average matrix of $X_{AP}$ along channel dimension, copy it along channel dimension by $C_{out} - C_{in}$, then concatenate the copied average matrix to $X_{AP}$. Tab. 1 shows the experimental results of shortcut candidates with three BNN models of which $C$, $k$, and $e$ are randomly selected. In all models, 'Partially duplicate' method achieves the highest accuracy among all candidates, thus we chose 'Partially duplicate' type as the shortcut path for the modified binary MBConv block.



(a) Zero

(b) Partially duplicate

(c) Global average

Figure 1: Candidates of shortcut path named (a) 'zero', (b) 'partially duplicate', and (c) 'global average' in a reduction binary MBConv block.

[*] Work done while at Pohang University of Science and Technology.

Table 1: Test accuracy of three randomly selected models R1 to R3 on CIFAR-100 dataset with various shortcut type for the reduction block. 'Z', 'PD', and 'GA' means 'zero', 'partially duplicate' and 'global average' shortcut type, respectively.

| Model | Equiv. Ops (M) | Test acc. (%) Z | PD | GA |
|---|---|---|---|---|
| R1 | 9.86 | 71.62 | **72.12** | 71.65 |
| R2 | 12.49 | 73.20 | **73.83** | 73.67 |
| R3 | 13.00 | 73.49 | **73.89** | 73.75 |

## 2. Analysis on search cost of BNN search with bimodal parameter perturbation

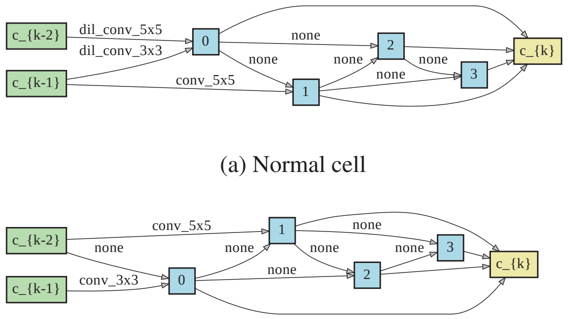
Tab. 2 shows the search and evaluation cost required to search for a model using our frameworks, which was measured with an RTX3090 GPU. Note that the noise injection process increases the search time by 20%. Grid search for the optimal noise scale point increases the total search cost (search time + training time), but the overhead was managed as follows. First, the range of noise scale is restricted as shown in Table 2 in the paper because larger noise prevents the search process from convergence. In case of ImageNet, previous works utilized the information obtained when searching for a model on smaller CIFAR dataset in the same search space as the proxy to reduce the heavy search cost [1]. We similarly searched for an optimal noise scale point in binary MBConv-based search space on CIFAR dataset, then reused the found noise scale to find a BNN model on ImageNet, and hence total search time was not increased by grid search in case of ImageNet.

Table 2: Search cost of our framework (with noise on both architecture and weight parameters) and average training cost of searched BNN models on each search space

| Dataset | Search Space | Search time (hours/GPU) | Training time (hours/GPU) |
|---|---|---|---|
| CIFAR-100 | BNAS | 9.8 | 21.4 |
| CIFAR-100 | MBConv | 26.5 | 26.7 |
| ImageNet-1K | MBConv | 73.3 | 853.3 |

## 3. Structures of searched cells on BNAS.

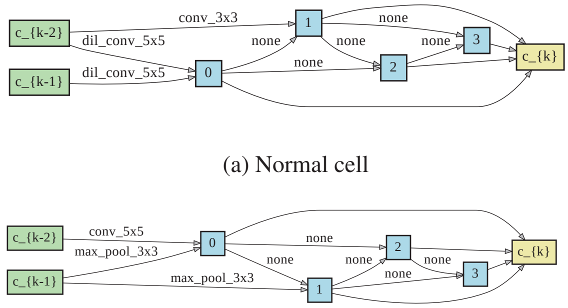
Figs. 2 to 5 illustrate the structures of searched cells on BNAS [2] under various perturbation conditions. As we compare Fig. 2 with Fig. 4, it is observed that injected noise on architecture parameters makes normal cell prefer pooling operations and reduction cell prefer convolutional operations. Perturbation on weight parameters regularizes the redundant operations: the number of operations in a cell is decreased and convolutional operations with smaller kernel



(a) Normal cell

(b) Reduction cell

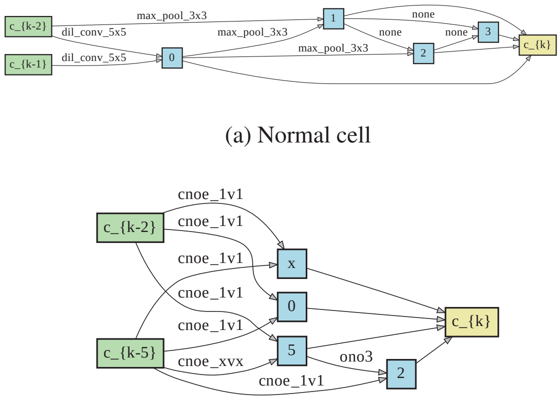
Figure 2: Searched (a) normal and (b) reduction cells on BNAS search space without parameter perturbation.



(a) Normal cell

(b) Reduction cell

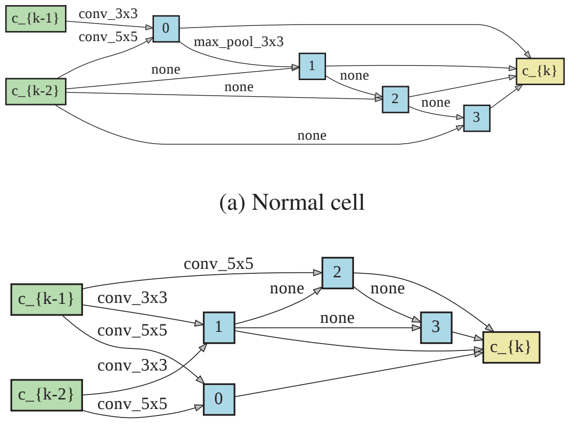
Figure 3: Searched (a) normal and (b) reduction cells on BNAS search space with perturbation on weight parameters ($\epsilon_W = 0.25$).



(a) Normal cell

(b) Reduction cell

Figure 4: Searched (a) normal and (b) reduction cells on BNAS search space with perturbation on architecture parameters ($\epsilon_A = 0.2$).



(a) Normal cell

(b) Reduction cell

Figure 5: Searched (a) normal and (b) reduction cells on BNAS search space with perturbation on weight and architecture parameters ($\epsilon_A = 0.2$ and $\epsilon_W = 0.25$).

size (e.g. 3) are preferred (Fig. 4 and Fig. 5).

## 4. Structures of searched models on binary MBConv block-based search space

Fig. 6 and Fig. 7 describe the structures of searched models on CIFAR and ImageNet dataset under various perturbation conditions. It is observed that injecting noise to weight parameters leads the search algorithm to prefer smaller kernel size (e.g. 3) and larger channel dimension of latter layers (Fig. 6). In case of the searched model on ImageNet dataset (Fig. 7), it is noticed that larger kernel size (e.g. 5) is preferred in the front layers and smaller kernel size (e.g. 3) is preferred in the latter layers. It seems that since the front layers process high-resolution features, a kernel with a larger reception field and representation capacity is preferred. On the contrary, the latter layers which process low-resolution features as input seem to prefer a small-sized kernel.

## 5. Comparison with previous works on CIFAR-10/100 datset

We compare our searched models on CIFAR-10/100 with the previous works in Tab. 3. Comparing the CIFAR-10/100 accuracy of our models on BNAS with bimodal perturbation and that of previous BNAS models, our models show similar accuracy with much smaller number of equivalent operations. With our frameworks on binary MBConv-based search space, our searched model (MBConv-A) can achieve higher accuracy on CIFAR-10/100 (93.76%, 73.61%) with similar number of equivalent operations (∼10M) compared with AresB-18 [3]. Our larger BNN models (MBConv-B, C, and D) shows much higher accuracy up to (94.38%, 75.53%) on CIFAR-10/100, respectively.

## References

[1] Xiangning Chen and Cho-Jui Hsieh. Stabilizing differentiable architecture search via perturbation-based regularization. In *International Conference on Machine Learning*, pages 1554–1565. PMLR, 2020.

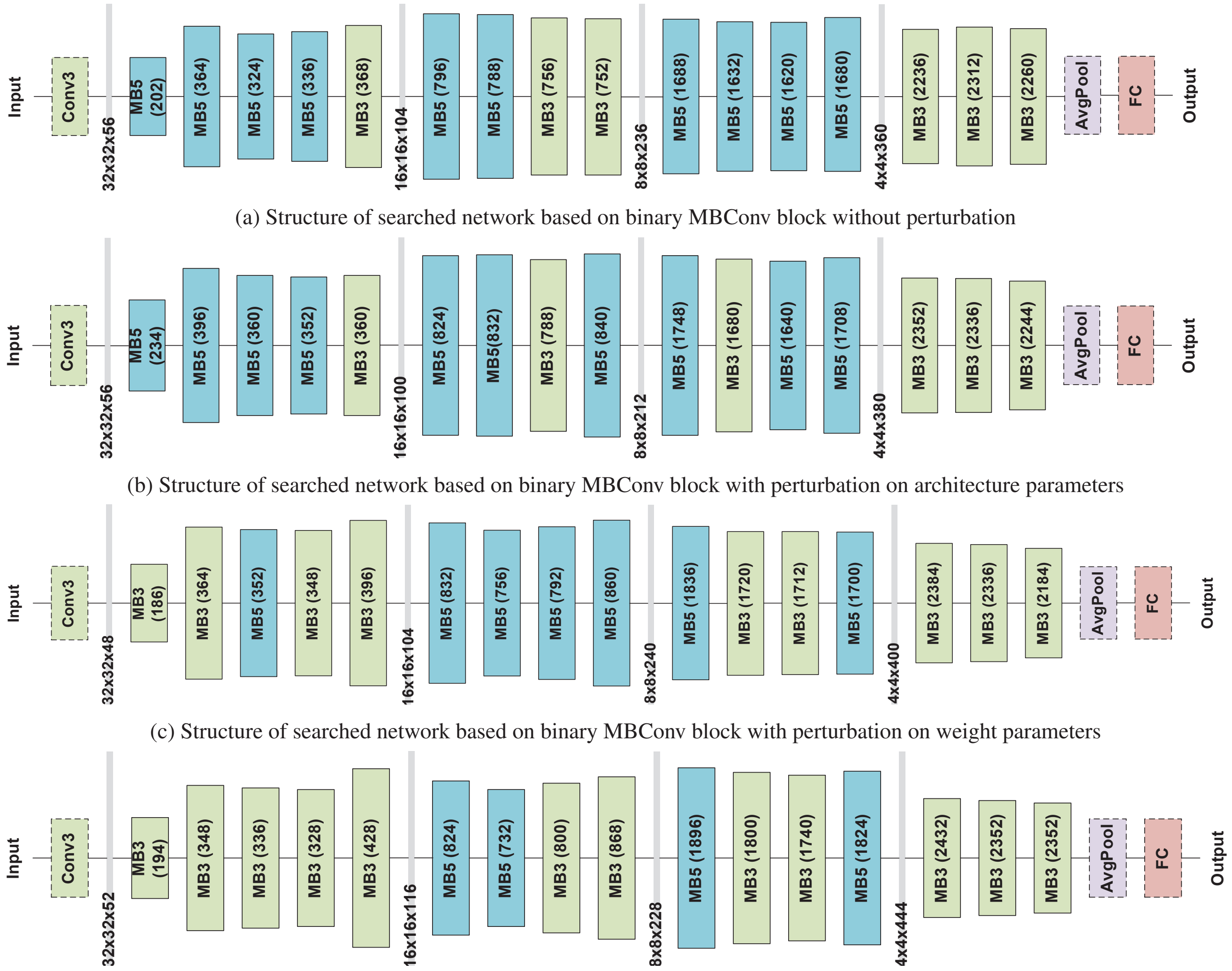
[2] Dahyun Kim, Kunal Pratap Singh, and Jonghyun Choi. Bnas v2: Learning architectures for binary networks with empirical improvements. *CoRR*, abs/2110.08562, 2021.

[3] HyunJin Kim. Aresb-net: accurate residual binarized neural networks using shortcut concatenation and shuffled grouped convolution. *PeerJ Computer Science*, 7:e454, 2021.

[4] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proceedings of the European conference on computer vision (ECCV)*, pages 722–737, 2018.

[5] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. *CoRR*, abs/1603.05279, 2016.

[6] Tianchen Zhao, Xuefei Ning, Xiangsheng Shi, Songyi Yang, Shuang Liang, Peng Lei, Jianfei Chen, Huazhong Yang, and Yu Wang. Bars: Joint search of cell topology and layout for accurate and efficient binary architectures, 2021.



(a) Structure of searched network based on binary MBConv block without perturbation

(b) Structure of searched network based on binary MBConv block with perturbation on architecture parameters

(c) Structure of searched network based on binary MBConv block with perturbation on weight parameters

(d) Structure of searched network based on binary MBConv block with perturbation on architecture and weight parameters

Figure 6: Structures of searched networks for CIFAR dataset based on binary MBConv block under diverse perturbation condition. ‘MB$k$ ($e$)’ stands for a binary MBConv block with binary $k \times k$ group ($g$=2) convolution and $e$ expanded channels. Each MBConv block in front of each gray line is a downsampling layer with stride=2, and the size of its output activation is changed as marked in the gray line. Real-valued operations are depicted by boxes with a dotted line.

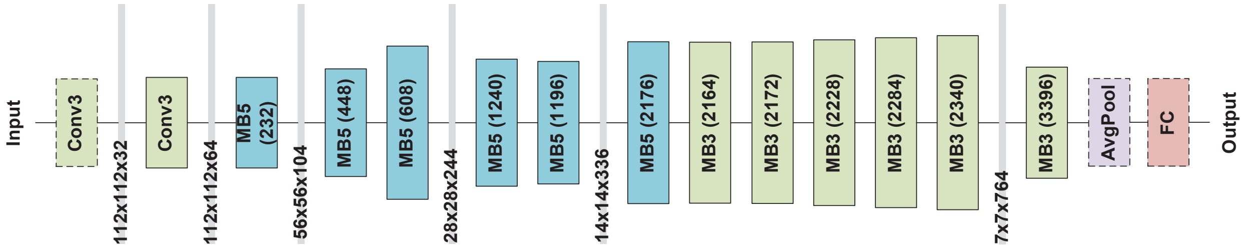


Figure 7: Structure of searched network based on binary MBConv block with perturbation on architecture and weight parameters on ImageNet dataset.

Table 3: Test accuracy on CIFAR10/100 dataset for various BNN models. XNOR-Net and Bi-Real-Net are based on ResNet-18. Ops and Params are calculated based on models for CIFAR-10 dataset, and in case of BNAS [2], Ops and Params are calculated including real-valued convolutions at shortcut paths in reduction blocks. 'Ours (BNAS-(mini, A, C$_m$))' models consist of (10,20,16) layers with (24,32,48) initial channels, respectively.

| Model | Search type | BOPs (G) | FLOPs (M) | Equiv. Ops (M) | Params (MB) | CIFAR-10 Acc. (%) | CIFAR-100 Acc. (%) |
|---|---|---|---|---|---|---|---|
| XNOR-Net [5] | Manual | 0.554 | 1.77 | 10.43 | 1.44 | 90 55 | N/A |
| Bi-Real-Net [4] | Manual | 0.547 | 8.07 | 16.62 | 2.11 | 91.23 | N/A |
| AresB-10 [3] | Manual | 0.245 | 1.77 | 5.608 | 0.63 | 90.74 | 69.45 |
| AresB-18 [3] | Manual | 0.547 | 1.77 | 10.33 | 1.43 | 91.90 | 73.01 |
| AresB-34 [3] | Manual | 0.849 | 1.77 | 15.05 | 2.23 | 92.71 | 74.73 |
| BNAS-A [2] | Gradient | 0.697 | 87.98 | 98.86 | 4.10 | 92.70 | N/A |
| BNAS-B [2] | Gradient | 1.330 | 273.8 | 294.6 | 11.7 | 93.76 | N/A |
| BNAS-C [2] | Gradient | 5.028 | 773.5 | 852.1 | 34.7 | 94.43 | N/A |
| BARS-A [6] | Gradient | 0.513 | 2.0 | 10.02 | N/A | 91.25 | N/A |
| BARS-B [6] | Gradient | 1.048 | 2.0 | 18.37 | N/A | 92.98 | N/A |
| BARS-C [6] | Gradient | 1.778 | 3.0 | 32.27 | N/A | 93.43 | N/A |
| **Ours (BNAS-mini)** | Gradient | 0.343 | 39.78 | 45.14 | 1.89 | **92.48** | **71.71** |
| **Ours (BNAS-A)** | Gradient | 1.326 | 87.98 | 108.7 | 4.75 | **94.21** | **75.38** |
| **Ours (BNAS-C$_m$)** | Gradient | 1.961 | 155.1 | 185.7 | 8.00 | **95.07** | **76.26** |
| **Ours (MBConv-A)** | Gradient | **0.567** | **1.36** | **10.58** | 1.94 | **93.76** | **73.61** |
| **Ours (MBConv-B)** | Gradient | **0.703** | **1.48** | **12.47** | 2.31 | **93.88** | **74.23** |
| **Ours (MBConv-C)** | Gradient | **0.821** | **1.60** | **14.43** | 2.64 | **94.20** | **74.72** |
| **Ours (MBConv-D)** | Gradient | **1.064** | **1.83** | **18.46** | 3.30 | **94.38** | **75.53** |