ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

A Key-Layer Selection-Based Bit-Flip Attack for Efficient Degradation of Quantized Neural Networks

Yiqing Wang[1,a], Jianqiang Mei[1,b,*], Fan Jia[2,c], Weixiang Du[3,d]

[1]*School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, China*
[2]*Raysov Instrument Co. Ltd., Dandong, China*
[3]*Gansu Province Special Equipment Inspection & Testing Research Institute, Lanzhou, Gansu, China*
[a]*2097455415@qq.com*, [b]*meijianqiang@tute.edu.cn*, [c]*info@raysov.com*, [d]*119273184@qq.com*
**Corresponding author*

Abstract: *Quantized neural networks (QNNs) are widely deployed on edge devices because of their high computational efficiency and low memory footprint. However, they remain vulnerable to hardware-level bit-flip attacks that can severely degrade model performance. Existing progressive bit-flip attack (PBFA) typically performs iterative searches across all network layers, resulting in substantial computational overhead and long attack time. To address this issue, we propose KS-BFA, a key-layer selection based bit-flip attack framework for quantized neural networks. The proposed method first evaluates the sensitivity of each layer and selects only the most vulnerable layers as attack targets, thereby avoiding exhaustive traversal across the entire network. Bit-flip attacks are then conducted only within these selected key layers, significantly reducing the search space and improving attack efficiency. Extensive experiments on CIFAR-10 and ImageNet demonstrate that KS-BFA can effectively degrade a variety of quantized models with only a small number of bit flips. For example, on ResNet-50, flipping about 22 bits reduces the model accuracy from 75.84% to 0.16%. Compared with PBFA, KS-BFA significantly reduces attack time while maintaining competitive destructive capability. In particular, on ResNet-50, the attack time decreases from 1330.83s to 1039.28s, corresponding to an efficiency improvement of approximately 22%. These results demonstrate that restricting bit-flip attacks to a small number of sensitive layers provides an efficient and scalable strategy for evaluating the vulnerability of quantized neural networks against hardware-aware bit-level attacks.*

Keywords: *Bit-Flip Attack, Key-Layer Selection, Quantized Neural Networks, Model Security, Attack Efficiency*

1. Introduction

Deep neural networks (DNNs) [1] have transformed numerous fields such as autonomous driving [2], medical diagnostics [3], and network security [4], enabling unprecedented levels of automation and precision. However, the high computational and memory requirements of full-precision [5] models limit their deployment on resource-constrained platforms such as edge devices [6] and embedded systems [7]. To address this issue, quantized neural networks (QNNs) [8] have been widely adopted because they significantly reduce model size and computational complexity by representing weights and activations with low-precision integers. Despite these advantages, quantized neural networks introduce new security vulnerabilities at the parameter level. Among these, Bit-Flip Attacks (BFAs) [9] have become increasingly alarming due to their subtle yet destructive impact: attackers manipulate the underlying model parameters by flipping specific bits stored in hardware memory [10], causing severe degradation in model accuracy with minimal modifications.

Quantized neural networks (QNNs) have been widely deployed on edge devices due to their high computational efficiency and low memory footprint. However, these networks remain vulnerable to hardware-level bit-flip attacks that can significantly degrade model performance. Existing methods, such as the Progressive Bit-Flip Attack (PBFA) [11], typically perform iterative searches across all network layers to identify vulnerable bits. While effective in degrading model accuracy, PBFA suffers from high computational overhead and long attack time, especially for deep neural networks with many layers.

To address these limitations, we propose KS-BFA, a Key-layer Selection Based Bit-Flip Attack framework. Instead of exhaustively searching across all layers, KS-BFA first evaluates the sensitivity of each layer and selects only the most vulnerable layers as attack targets. Bit-flip attacks are then conducted

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

solely within these key layers, substantially reducing the search space and improving attack efficiency. Extensive experiments on CIFAR-10 and ImageNet demonstrate that KS-BFA can effectively degrade the performance of quantized models with only a small number of bit flips. The main contributions of this paper are summarized as follows:

(1) We propose KS-BFA, a key-layer selection based bit-flip attack framework for quantized neural networks, which improves attack efficiency by restricting the attack to a small subset of sensitive layers.

(2) We directly perform bit-flip attacks on selected critical layers, thereby avoiding an exhaustive traversal of all network layers and significantly reducing the search space and computational overhead of the bit-flip attacks.

(3) Extensive experiments on CIFAR-10 and ImageNet demonstrate that KS-BFA achieves competitive destructive effectiveness while significantly reducing attack time compared with PBFA.

The remainder of this paper is organized as follows. Section 2 reviews related work, introduces the model quantization process, and describes the threat model considered in this study. Section 3 presents the proposed KS-BFA framework, including the key-layer selection strategy and the bit-flip attack procedure. Section 4 reports the experimental setup and evaluation results on CIFAR-10 and ImageNet, together with comparisons against existing attack methods. Finally, Section 5 concludes the paper.

2. Related Work

2.1 Parameter Attacks

As Parameter-level adversarial attacks are receiving increasing attention in the security analysis of deep neural networks [12]. Due to the highly nonlinear nature of DNNs, even small perturbations to input data or model parameters may lead to significant changes in prediction results [13]. Many adversarial weight attack methods aim to maximize the impact of the attack by identifying parameters with large gradients with respect to the inference loss and then introducing small perturbations to these weights [14]. Building upon these ideas, Bit-Flip Attack (BFA) methods further exploit the discrete representation of quantized neural network parameters. Instead of directly modifying floating-point weights, BFA manipulates the binary representation of quantized weights stored in memory [15] by flipping selected bits. By carefully identifying and flipping a small number of critical bits, attackers can significantly increase the inference loss and drastically degrade model accuracy. Compared with traditional adversarial weight perturbations, BFA is particularly practical in real-world scenarios because it directly targets the bit-level storage of quantized models, which are widely used in resource-constrained machine learning services [16].

2.2 Model Quantization

To evaluate the vulnerability of quantized neural networks under bit-flip attacks, we adopt a layer-wise uniform quantization strategy, which is widely used in previous studies on quantized model security. Given a pretrained full-precision neural network, each layer is independently quantized into an (N)-bit fixed-point representation before the attack is performed. Its specific definition is as follows:

$$\Delta w = \frac{\max\left(\mathrm{W}^{\mathrm{fp}}\right)}{2^{N-1}-1}; \quad \mathrm{W}^{\mathrm{fp}} \in \mathbb{R}^{d} \tag{1}$$

$$\mathrm{W} = \mathrm{round}\left(\frac{\mathrm{W}^{\mathrm{fp}}}{\Delta w}\right) \cdot \Delta w \tag{2}$$

where W^{fp} is the full-precision floating-point weight tensor (e.g., FP32), d is the dimension of the weight tensor, Δw is the step size of the weight quantizer, N is the quantization bit width, and W is the quantized weight.

2.3 Threat Model

This study considers a white-box attacker whose goal is to degrade the performance of a deployed quantized neural network by manipulating its bit-level parameters. The attacker is assumed to have access

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

to the model architecture and quantized parameter values, while the network structure and input data stream remain unchanged. The attack aims to flip only a very limited number of bits in the stored parameters so as to cause substantial degradation in model accuracy or even complete failure. Following prior studies on hardware-aware bit-flip attacks, we assume that the flips can be induced through memory fault mechanisms such as Rowhammer in DRAM-based systems [17].

3. Methodology

This section describes the proposed KS-BFA method in detail. The objective of KS-BFA is to efficiently identify vulnerable bits in quantized neural networks while reducing the search overhead of conventional progressive bit-flip attacks. As illustrated in Figure 1, the overall workflow consists of three stages: preprocessing and quantization of the clean full-precision model, sensitivity-based key-layer selection, and bit-flip attacks within the selected layers.

Specifically, the pretrained full-precision model is first converted into a quantized neural network through layer-wise quantization. Next, the sensitivity of each layer is estimated using the L2 norm of the loss gradient with respect to its binary weights, and the most sensitive layers are selected as candidate attack targets. Finally, vulnerable bits are iteratively identified and flipped within the selected key layers until the attack objective is achieved.

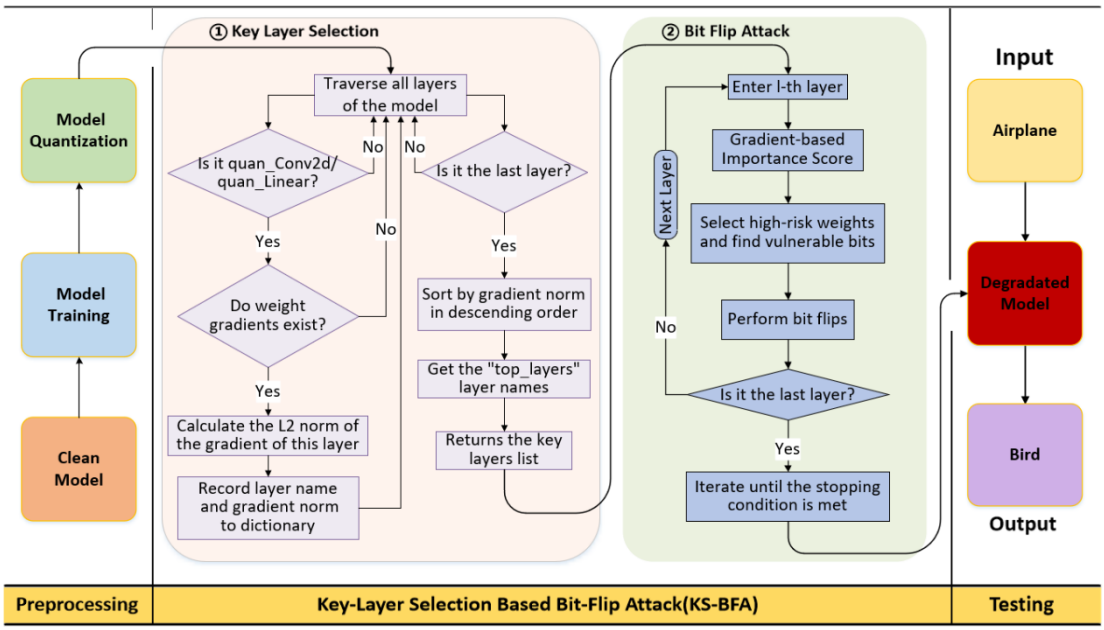

Figure 1: Overall workflow of the proposed KS-BFA framework.

3.1 Key Layer Selection

Directly searching for vulnerable bits across all layers of a quantized neural network is computationally expensive and often unnecessary, because different layers contribute unequally to the final prediction. In practice, some layers are substantially more sensitive to parameter perturbations than others. To improve attack efficiency, KS-BFA first identifies a small set of key layers and restricts the subsequent bit search to these layers.

We quantify layer sensitivity using the L2 norm of the loss gradient with respect to the binary weights of each layer. Intuitively, a larger gradient norm indicates that small perturbations to the corresponding layer are more likely to induce a larger increase in loss and, consequently, a larger change in model output. For each layer l, the layer importance score is defined as:

$$G_l = \left\| \nabla_{\mathrm{B}_l} \mathcal{L} \right\|_2 \tag{3}$$

where $\mathcal{L}$ denotes the classification loss and $\nabla_{\mathrm{B}_l}\mathcal{L}$ is the gradient of the loss with respect to the binary weight tensor B_l of layer l. A larger value of G_l indicates that layer l is more influential with respect to the attack objective.

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

After computing G_l for all layers $l \in \{1,2,\dots,L\}$, we rank the layers in descending order of G_l and select the top-k layers as the key attack layers:

$$L_{\text{key}} = \arg\text{Top}-k_{l\in\{1,\dots,L\}} G_l \tag{4}$$

By restricting the attack to the most sensitive layers, this strategy substantially reduces the search space and computational cost while preserving strong destructive effectiveness.

3.2 Bit-Flip Attack

After identifying the key layers, KS-BFA performs iterative bit-flip attacks only within the restricted search space. Unlike PBFA, which progressively traverses all layers during the attack process, the proposed method focuses exclusively on the most sensitive layers selected by the key-layer strategy. This significantly reduces computational overhead and attack time.

For each attacked layer $l \in L_{\text{key}}$, the importance of binary weights is estimated using the absolute value of the loss gradient with respect to the binary weight tensor:

$$I_l = \left|\nabla_{\text{B}_l}\mathcal{L}(f(x;\text{B}),t)\right|, \qquad l \in L_{\text{key}} \tag{5}$$

where x and t denote the input samples and their corresponding ground-truth labels, respectively. $\nabla_{\text{B}_l}\mathcal{L}(\cdot)$ denotes the gradient of the loss with respect to the binary weights of layer l.

For each attacked layer l, all candidate bits are ranked according to their importance scores, and the top n_b bits (n_b = 1 by default) are selected for flipping:

$$S_l = \arg\text{Top}_{n_b}(I_l), \qquad \widehat{\text{B}}_l[S_l] = \text{B}_l[S_l] \oplus m \tag{6}$$

where S_l denotes the index set of the n_b most important bits in layer l, m is a binary mask whose entries are set to 1 at the selected bit positions and 0 elsewhere, and $\oplus$ denotes the bitwise XOR operation. $\text{B}_l[S_l]$ represents the value of B_l at index S_l and $\widehat{\text{B}}_l[S_l]$ denotes the value of B_l at index S_l after the selected bits have been flipped.

4. Experiments

4.1 Experimental Setup

To evaluate the proposed KS-BFA comprehensively, we conduct experiments on multiple neural network architectures across the CIFAR-10 and ImageNet datasets. All experiments are performed on a Linux server equipped with an NVIDIA A800 GPU with 80 GB memory. The attack framework is implemented in Python 3.8 using the PyTorch 2.4 deep learning library.

4.2 Datasets and Network Architectures

To comprehensively evaluate the proposed attack strategy, we conduct experiments on two widely used image classification benchmarks: CIFAR-10 and ImageNet. CIFAR-10 contains 60,000 RGB images of size 32 × 32, including 50,000 training images and 10,000 test images, whereas ImageNet (ILSVRC-2012) contains approximately 1.2 million training images and 50,000 validation images spanning 1,000 categories. For attack generation, we randomly sample a small labeled subset from the training set to compute gradients and activation statistics. For post-attack evaluation, we randomly sample 128 images from the CIFAR-10 test set and 256 images from the ImageNet validation set to measure the degraded model accuracy. On CIFAR-10, we use VGG-11 and ResNet-20/32/56 as representative convolutional neural networks, all trained from scratch for 160 epochs using SGD with a learning rate of 0.1, a batch size of 128, and a weight decay of 0.0003. On ImageNet, we adopt the official pretrained models, including AlexNet and ResNet-18/34/50. In all experiments, the models are quantized to 8-bit weights before applying KS-BFA.

4.3 Experiments on CIFAR-10

We first evaluate the effectiveness of KS-BFA on the CIFAR-10 dataset using several widely used convolutional neural network architectures, including VGG-11, ResNet-20, ResNet-32, and ResNet-56. These models provide a representative set of shallow and moderately deep backbones for assessing the

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

attack under low-resolution image classification settings. All models are quantized to 8-bit weights before the attack is performed.

Table 1: Performance degradation of quantized models on CIFAR-10 under KS-BFA.

Model	Baseline acc.	Quantized acc.(%)	Post-Attack acc.(%)	# of Bits Flipped
VGG-11	92.30	91.06	10.89±0.32	45.5±15.41
ResNet-20	92.24	92.34	10.40±0.22	13±3.16
ResNet-32	92.92	93.28	10.89±0.08	21.2±7.05
ResNet-56	93.38	93.06	10.90±0.04	18.5±5.64

As shown in Table 1, KS-BFA successfully reduces the accuracy of all evaluated models to approximately 10%, which corresponds to random guessing for a ten-class classification task. These results indicate that severe degradation can be achieved with only a limited number of bit flips. In particular, ResNet-20 requires only 13±3.16 bit flips to reduce the classification accuracy from 92.34% to 10.40%, while ResNet-56 requires only 18.5±5.64 bit flips. Similar trends can also be observed for VGG-11 and ResNet-32.

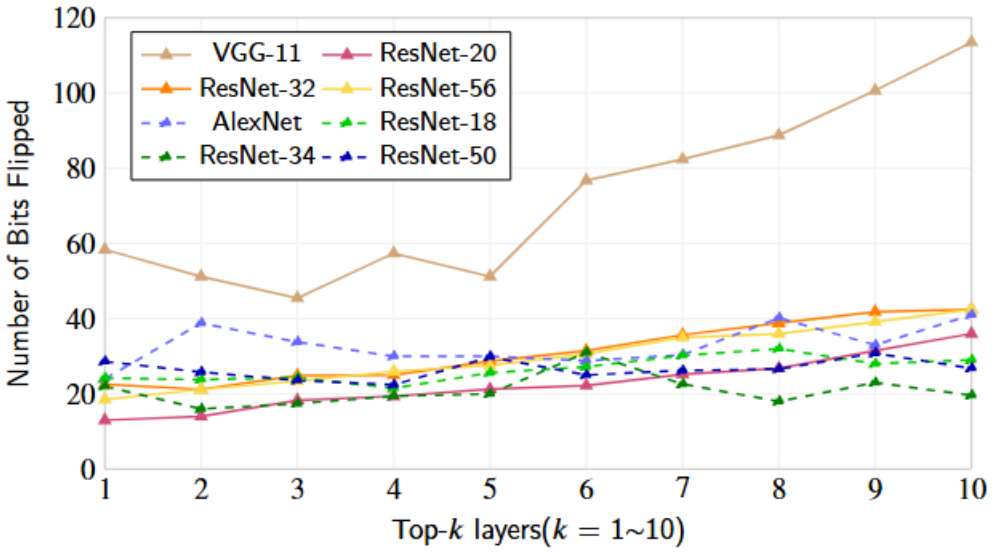

Figure 2: Effect of the number of key layers (k=1~10) on the number of flipped bits required by KS-BFA across multiple neural network architectures on CIFAR-10 and ImageNet.

To further analyze the effect of the key-layer selection strategy, Figure 2 shows the relationship between the number of selected key layers and the number of bit flips required for a successful attack. In general, the required number of bit flips increases as more layers are included in the search space. This trend is especially pronounced for VGG-11, whose curve rises more steeply than those of the ResNet models. These observations indicate that restricting the attack to a small set of highly sensitive layers is beneficial for improving attack efficiency.

These results demonstrate that restricting attacks to a small subset of sensitive layers does not compromise attack effectiveness. Instead, KS-BFA remains capable of identifying highly vulnerable parameters and causing catastrophic performance degradation with only a few bit modifications.

4.4 Experiments on ImageNet

To further evaluate the scalability and effectiveness of the proposed attack, we conduct experiments on the ImageNet dataset, which presents a considerably more challenging setting than CIFAR-10 because of its larger number of classes and higher-resolution images. The evaluated architectures include AlexNet, ResNet-18, ResNet-34, and ResNet-50. As in the CIFAR-10 experiments, all models are quantized to 8-bit weights before the attack is applied.

The results are presented in Table 2. Similar to the observations on CIFAR-10, KS-BFA remains highly effective on large-scale image classification tasks. For all tested architectures, model accuracy is reduced from over 56% – 76% to less than 0.2%. For example, ResNet-50 achieves 75.84% top-1 accuracy after quantization, but its accuracy drops to only 0.16% after approximately 22 bit flips.

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

Likewise, ResNet-34 is reduced from 73.13% to 0.16% with only 16 flipped bits.

Table 2: Performance degradation of quantized models on ImageNet under KS-BFA.

Model	Baseline acc.	Quantized acc.(%)	Post-Attack acc.(%)	# of Bits Flipped
AlexNet	56.55	56.16	0.19±0.01	24±10.46
ResNet-18	69.64	69.49	0.19±0.01	21.5±6.28
ResNet-34	73.29	73.13	0.16±0.02	16±3.56
ResNet-50	76.15	75.84	0.16±0.01	22.4±6.83

Similarly, on ImageNet, Figure 2 illustrates the relationship between the number of selected key layers and the number of bit flips required for the model. Across all tested architectures, even as the search space expands, the required number of bit flips remains remarkably stable, typically fluctuating between 10 and 40 bits. Unlike the trend observed on smaller datasets, models on large-scale datasets exhibit more evenly distributed vulnerabilities across their key layers, with AlexNet showing the largest variation in required bit flips. These results further confirm that focusing on a small subset of highly sensitive layers is sufficient to compromise the performance of quantized models on large-scale tasks.

To improve the reliability of the results reported in Table 1 and Table 2, each key-layer configuration (i.e., each Top-k setting) is evaluated over 5 independent trials. For each architecture, the reported number of bit flips corresponds to the minimum number of flips required to reach the target accuracy degradation under a given configuration, and the final results are reported as the mean $\pm$ standard deviation over repeated runs. This protocol reduces the influence of randomness in bit selection and improves the reproducibility of the reported attack performance.

4.5 Comparison with PBFA

To evaluate the efficiency of the proposed key-layer selection strategy, KS-BFA is compared with the representative Progressive Bit-Flip Attack (PBFA). The comparison results are reported in Table 3.

Table 3: Comparison of KS-BFA and PBFA across datasets and network architectures.

Dataset	Model	KS-BFA		PBFA	
		N_{flip}	Attack times (s)	N_{flip}	Attack times (s)
CIFAR-10	VGG-11	45.5±15.41	88.79±35.82	38.6±24.95	84.55±42.08
	ResNet-20	13±3.16	25.56±13.23	12.8±5.36	31.43±13.15
	ResNet-32	21.2±7.05	44.02±12.71	20±11.03	50.97±27.25
	ResNet-56	18.5±5.64	49.01±21.07	17.6±6.40	56.21±20.70
ImageNet	AlexNet	24±10.46	2382.24±946.67	19.5±4.22	2435.06±515.81
	ResNet-18	21.5±6.28	1241.69±299.94	13.5±1.43	1269.78±129.68
	ResNet-34	16±3.56	902.77±208.54	10±3.62	942.68±355.22
	ResNet-50	22.4±6.83	1039.28±320.66	17±2.57	1330.83±244.16

As shown in Table 3, KS-BFA consistently reduces attack time across all evaluated architectures while maintaining competitive attack effectiveness. On CIFAR-10, the attack time reduction ranges from approximately 10% to 20%. On ImageNet, the improvement becomes more pronounced due to the larger network depth and search space. For example, on ResNet-50, KS-BFA completes the attack in 1039.28s, whereas PBFA requires 1330.83s. This corresponds to an efficiency improvement of approximately 22%. Similar improvements can be observed on ResNet-18, ResNet-34, and AlexNet. Although PBFA occasionally requires slightly fewer bit flips, it performs exhaustive searches across all layers. By contrast, KS-BFA focuses only on the most sensitive layers, substantially reducing computational overhead while preserving comparable destructive capability.

These results demonstrate that key-layer selection provides an effective means of accelerating bit-flip attacks without significantly sacrificing attack effectiveness.

4.6 Comparison with Random Bit-Flips

We further investigate the impact of random bit flips compared to KS-BFA. We analyze the robustness of the 8-bit quantized ResNet-50 model against random bit flips. As illustrated in Figure 3, flipping up to 100 random bits results in less than a 2% drop in both Top-1 and Top-5 accuracies on ImageNet. This underscores the necessity for targeted bit-flip strategies, as random flips alone fail to significantly impair model performance. In contrast, our proposed KS-BFA completely disrupts ResNet-50 classification accuracy on ImageNet by flipping only 22 strategically selected bits out of 204 million.

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

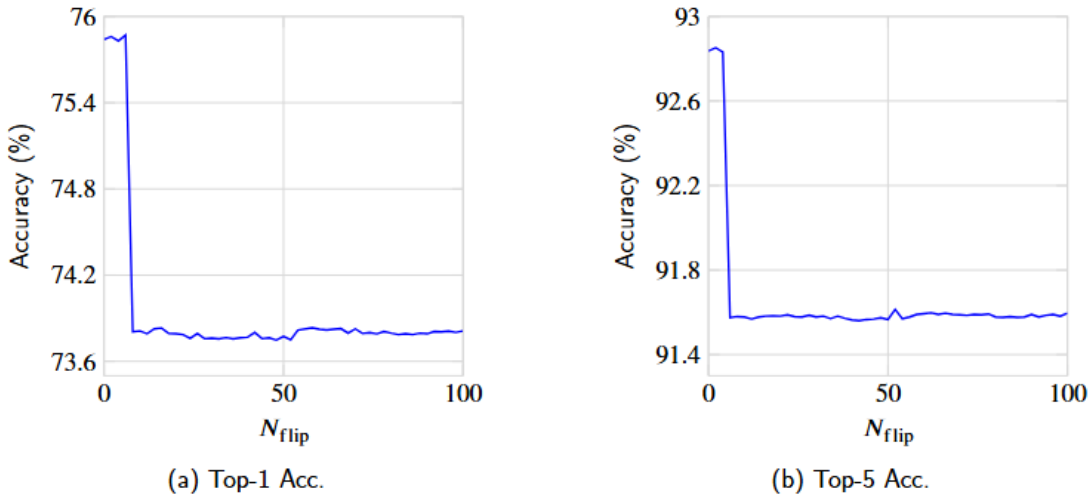

Figure 3: Effect of randomly flipping 100 bits in ResNet-50 on ImageNet, measured by Top-1 and Top-5 accuracy.

5. Conclusion

This paper presented KS-BFA, a key-layer selection based bit-flip attack framework for quantized neural networks. Unlike conventional progressive bit-flip attacks that traverse all network layers, the proposed method first identifies the most sensitive layers and restricts the attack to these layers only. This strategy substantially reduces the search space and improves attack efficiency. Extensive experiments on CIFAR-10 and ImageNet demonstrated that KS-BFA can effectively degrade the performance of various quantized neural networks with only a small number of bit flips. Compared with PBFA, KS-BFA significantly reduces attack time while maintaining competitive destructive effectiveness. In particular, on ResNet-50, the attack time is reduced from 1330.83s to 1039.28s, corresponding to an efficiency improvement of approximately 22%. These findings indicate that focusing bit-flip attacks on sensitive layers provides an efficient and scalable strategy for evaluating the vulnerability of quantized neural networks under hardware-level attacks.

References

[1] Djeddou M, Dallal J A, Hellal A, et al. Particle Swarm Optimization-Based Deep Neural Network vs Whale Optimization Algorithm-Based Deep Convolutional Neural Networks for Critical Heat Flux Prediction[C]//2024 International Conference on Decision Aid Sciences and Applications (DASA). Manama, Bahrain, 2024: 1-5.

[2] Cheng K, Zhou Y, Chen B, et al. Guardauto: A Decentralized Runtime Protection System for Autonomous Driving[J]. IEEE Transactions on Computers, 2021, 70(10): 1569-1581.

[3] Singh D, Verma S, Singla J. A Comprehensive Review of Intelligent Medical Diagnostic Systems[C]//2020 4th International Conference on Trends in Electronics and Informatics (ICOEI). Tirunelveli, India, 2020: 977-981.

[4] Lv Z, Chen D, Cao B, et al. Secure Deep Learning in Defense in Deep-Learning-as-a-Service Computing Systems in Digital Twins[J]. IEEE Transactions on Computers, 2024, 73(3): 656-668.

[5] Lumoindong C W D, Mandala R. Binarized and Full-Precision 3D-CNN in Action Recognition[C]//2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE). Yogyakarta, Indonesia, 2022: 241-246.

[6] Benbraika M K, Bourekkache S, Kraa O, et al. Mobile Edge Computing Discovery for Device-to-Device Communication in 5G[C]//2024 8th International Conference on Image and Signal Processing and their Applications (ISPA). Biskra, Algeria, 2024: 1-5.

[7] Islas-Estrada L R, Flores-Hernández D A. Embedded Energy Monitoring System for Solar Applications[J]. IEEE Embedded Systems Letters, 2025, 17(6): 386-389.

[8] Surapally S K, Yang X, Harman T L, et al. Evaluating FPGA Acceleration on Binarized Neural Networks and Quantized Neural Networks[C]//2022 International Symposium on Measurement and Control in Robotics (ISMCR). Houston, TX, USA, 2022: 1-5.

[9] Hector K, Moëllic P A, Dumont M, et al. A Closer Look at Evaluating the Bit-Flip Attack Against

ISSN 2616-5775 Vol. 9, Issue 6: 29-36, DOI: 10.25236/AJCIS.2026.090605

Deep Neural Networks[C]//2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS). Torino, Italy, 2022: 1-5.

[10] Zhang Z, et al. Implicit Hammer: Cross-Privilege-Boundary Rowhammer Through Implicit Accesses[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(5): 3716-3733.

[11] Rakin A S, He Z, Fan D. Bit-Flip Attack: Crushing Neural Network With Progressive Bit Search[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), 2019: 1211-1220.

[12] Irfan M M, Ali S, Yaqoob I, et al. Towards Deep Learning: A Review On Adversarial Attacks[C]//2021 International Conference on Artificial Intelligence (ICAI). Islamabad, Pakistan, 2021: 91-96.

[13] Delattre B. On the Stability of Neural Networks in Deep Learning[J/OL]. arXiv preprint, 2025.

[14] Wu D, Xia S-T, Wang Y. Adversarial Weight Perturbation Helps Robust Generalization[C]//Advances in Neural Information Processing Systems. Curran Associates, Inc., 2020: 2958-2969.

[15] Bai J, Wu B, Zhang Y, et al. Targeted Attack against Deep Neural Networks via Flipping Limited Weight Bits[J/OL]. arXiv preprint, 2021.

[16] Abomakhelb A, Jalil K A, Buja A G, et al. A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks[J]. Technologies, 2025, 13(5): 202.

[17] Cheng Q, Ming Z, Yuanping N, et al. A Survey of Bit-Flip Attacks on Deep Neural Network and Corresponding Defense Methods[J]. Electronics, 2023, 12(4): 853.