ZeBRA: Precisely Destroying Neural Networks with Zero-Data Based Repeated Bit Flip Attack (Supplementary Material)

Dahoon Park*[1]
pdh930105@dgist.ac.kr
Kon-Woo Kwon[2]
konwoo@hongik.ac.kr
Sunghoon Im[1]
sunghoonim@dgist.ac.kr
Jaeha Kung**[1]
jhkung@dgist.ac.kr

[1] Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea
[2] Hongik University, Seoul, Korea

A Attack Performance of ZeBRA at Different Batch Sizes

Table 1 summarizes the impact of the mini-batch size $B_{distill}$ on the ZeBRA attack performance. With $B_{distill} = 16$, only 8bits are required to be flipped to achieve 8.72% of accuracy on ResNet-20 with CIFAR-10 dataset. It showed similar results on the other benchmarks and DNN models providing $B_{distill} = 16$ to be the best choice for the ZeBRA attack. Thus, we selected $B_{distill} = 16$, $B_{attack} = 64$, $\lambda_{CE} = 0.2$, $\lambda_{Distill} = 0.1$, and $\varepsilon_{loss} = 10$ when generating $\mathbf{X}^{d}_{attack}$ for the entire experiments.

Table 1: The attack performance of ZeBRA at various distill batch sizes of the distilled target data (B_{attack} = 64, $N_b^{\max}$ = 40)

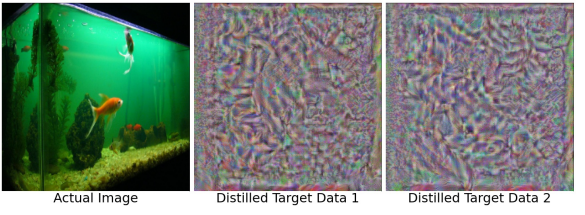
Distill Batch Size ($B_{distill}$)	Bit Flips (Best)	Bit Flips (Mean)	Bit Flips (Worst)	Acc. [%] (Best)	Acc. [%] (Mean)
8	8	14.78	30	8.77	9.93
16	**8**	**11.56**	33	8.72	9.58
32	17	33.40	40	9.99	10.02
64	19	35.98	40	10.00	10.01

B Visualization of Distilled Target Data

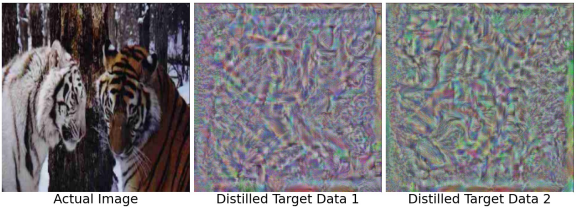
Figure 1 compares the actual data and the distilled target data synthesized by setting $\{\lambda_{CE} = 0.2, \lambda_{Distill} = 0.1, \varepsilon_{loss} = 10, B_{distill} = 16\}$ on ResNet-18. We provide two distilled target data for each image sample trained by the ZeBRA algorithm. Although synthesized data are visually far from the real data, they preserve the statistics of batch normalization layers for

*D. Park was with Hongik University when started working on this paper (**Corresponding author: J. Kung).

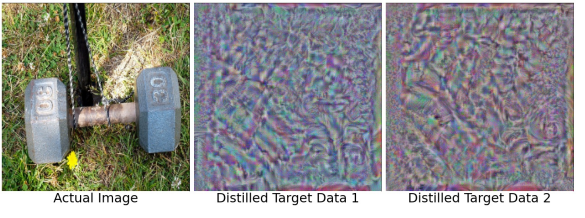
more efficient attack. The visualization of the input features to the last layer via t-SNE [8] shows that the real and distilled target data of the same target label are clustered nearby in the projected feature space by adding the distilled loss ($\lambda_{Distill} = 0.1$; right of Figure 2) when generating the data. With $\lambda_{Distill} = 0.0$ (left of Figure 2), the distilled target data on some target labels do not cluster near the real data. It empirically implies that utilizing only the cross-entropy loss is not sufficient to guide the distilled target data to be projected in the desired feature space at the classifier. The effectiveness of the ZeBRA attack on various λ_{CE} and $\lambda_{Distill}$ is reported in Section 4.1.



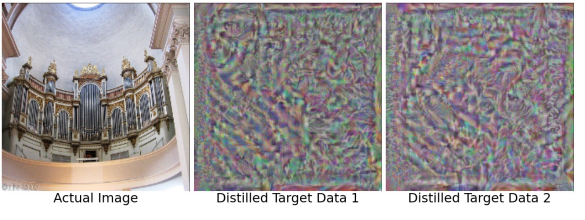
(a) Gold Fish



(b) Tiger



(c) Dumbbell



(d) Organ

Figure 1: Comparison between the actual data and distilled target data ($\lambda_{CE} = 0.2$, $\lambda_{Distill} = 0.1$, $\varepsilon_{loss} = 10$, $B_{distill} = 16$) tested on ResNet-18.

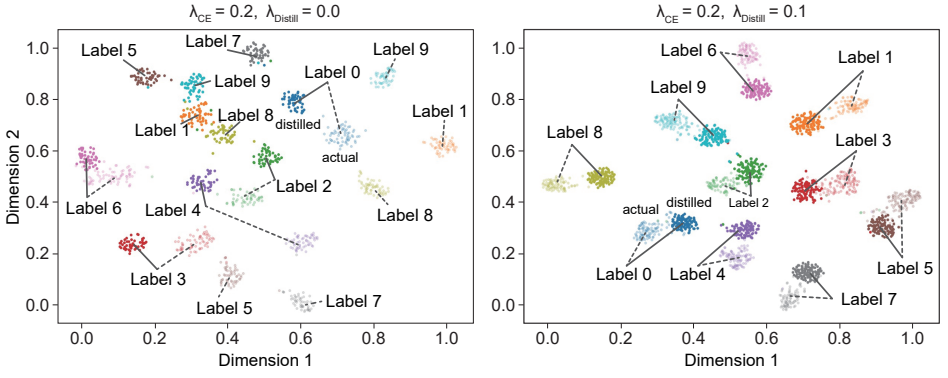


Figure 2: The t-SNE of the actual dataset and the distilled target data at the classifier of ResNet-20 on CIFAR-10 with different hyper-parameters: (left) $\lambda_{CE} = 0.2$, $\lambda_{Distill} = 0.0$, $\varepsilon_{loss} = 10$ and (right) $\lambda_{CE} = 0.2$, $\lambda_{Distill} = 0.1$, $\varepsilon_{loss} = 10$.

Table 2: The comparison of the attack performance on mobile-friendly DNN models between the BFA and the proposed ZeBRA on ImageNet dataset

Model	Accuracy	Original Accuracy [%]	BFA				ZeBRA			
			Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Avg. Accuracy After Attack	Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Avg. Accuracy After Attack
MobileNetV2	**Top-1**	71.14	1	8	2.65 / 0.031	0.14 0.66	1	2	1.68 / 0.014	0.12 0.61
	Top-5	90.01								
ShuffleNetV2 0.5	**Top-1**	59.36	1	4	2.69 / 0.62	0.13 1.01	1	3	2.60 / 0.48	0.12 0.91
	Top-5	81.07								
ShuffleNetV2 1.0	**Top-1**	68.68	2	5	3.60 / 0.80	0.13 0.64	3	3	3.00 / 0.00	0.16 0.77
	Top-5	87.92								
MnasNet 0.5	**Top-1**	66.31	1	5	2.25 / 0.99	0.12 0.96	1	3	1.80 / 0.68	0.12 0.89
	Top-5	86.76								
MnasNet 1.0	**Top-1**	72.34	1	3	1.60 / 0.66	0.12 0.90	1	2	1.10 / 0.30	0.12 1.02
	Top-5	90.73								

C Evaluation on Mobile-friendly DNN Models

As discussed in Section 4.2.2, a mobile-friendly model can be significantly damaged by the adversarial weight attack due to the existence of pointwise convolutions. To support the statement that mobile-friendly DNN models are more vulnerable, we evaluated the attack performance of the ZeBRA on MobileNetV2 [5], ShuffleNetV2 [4], and MnasNet [7]. As summarized in Table 2, the ZeBRA attack on most mobile-friendly DNN models results in <0.2% Top-1 accuracy even by flipping one bit (3 bits are required for ShuffleNetV2 1.0). Table 3 shows the attack performance of the ZeBRA when evaluated by the distilled validation data as discussed in Section 4.3.

D Evaluation of ZeBRA on Additional Datasets

We conducted additional experiments to show that DNN models trained on medical datasets are also vulnerable to ZeBRA attack. First, brain tumor dataset [2] consists of 2D MRI slice images and there exist three different types of brain tumors to be classified. We use the same pre-processing on training data as presented in [6]. The tested DNN models are ResNet-18 and MobileNetV2. Compared to the BFA attack, the ZeBRA destroys the trained model on

Table 3: The attack performance of the ZeBRA when evaluated by the distilled validation data on mobile-friendly DNN models

Model	ZeBRA w/ Distilled Validation Data (ImageNet)				
	Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Top-1 Accuracy (Min / Avg)	Top-5 Accuracy (Min / Avg)
MobileNetV2	1	4	2.40 / 0.80	0.10 / 1.10	0.84 / 5.70
ShuffleNetV2 0.5	1	3	1.85 / 0.65	0.11 / 1.38	0.64 / 3.85
ShuffleNetV2 1.0	2	3	2.21 / 0.41	0.15 / 2.71	0.62 / 6.70
MnasNetV2 0.5	2	3	2.05 / 0.21	0.092 / 1.28	0.46 / 4.43
MnasNetV2 1.0	2	3	2.27 / 0.44	0.10 / 7.57	0.83 / 11.47

Table 4: The comparison of the attack performance between the BFA and the proposed ZeBRA on the brain tumor dataset ($B_{attack} = 32$, $A_{target} = 35\%$)

Model	Original Accuracy [%]	BFA				ZeBRA			
		Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Avg. Accuracy After Attack	Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Avg. Accuracy After Attack
ResNet-18	98.58	4	9	6.6 / 1.66	33.85	4	8	6.3 / 0.71	34.02
MobileNetV2	97.87	1	3	1.66 / 0.66	33.53	1	2	1.33 / 0.42	33.59

the brain tumor dataset with 7.6% less number of bit flips on average (Table 4). Here, the target accuracy is set to 35% which means that the model has turned into a random predictor. Another medical dataset under test is a skin disease dataset called DermNet dataset [1]. It contains 23 classes and 15,557 training images and 4,002 test images. We trained ResNet-50 and MobileNetV2 on the DermNet dataset and performed both the BFA and ZeBRA attacks for the comparison. As a result, the ZeBRA successfully destroys DNN models with 22.5% less number bit flips on average (Table 5). According to this additional evaluation on two medical datasets, we can generalize the effectiveness of the ZeBRA attack on a wide range of applications.

E ZeBRA on Fortified DNN Models

In [3], authors found that the BFA tends to select and flip MSBs making small weights large values (e.g., from 0 to ± 128). The same observations were made when we perform the ZeBRA on all DNNs. Based on this observation, piece-wise clustering of weight parameters has been proposed to avoid the large shift in weight values [3]. To perform the piece-wise clustering, the following penalty term is added to CE loss during the training:

$$\lambda_{cluster} \cdot \sum_{l=0}^{L-1} (||\theta_l^+ - \mathbb{E}(\theta_l^+)||_2 + ||\theta_l^- - \mathbb{E}(\theta_l^-)||_2), \tag{1}$$

where $\lambda_{cluster}$ is the cluster coefficient and θ_l^+ (or θ_l^-) denotes the positive (or negative) weight parameters at l-th layer. By adding Eq. (1), the weight distribution of a DNN model changes from Gaussian to bimodal distribution.

We call the DNN model with piece-wise clustering 'the fortified DNN model'. After training the fortified ResNet-20 on CIFAR-10, we compared the attack performance between the BFA and ZeBRA at various quantization levels (N_Q) and cluster coefficients ($\lambda_{cluster}$). The Top-1 accuracy prior to the adversarial weight attack is reported in Table 6. As shown in Figure 3, the average number of bit flips to reach 10% Top-1 accuracy using the ZeBRA

Table 5: The comparison of the attack performance between the BFA and the proposed ZeBRA on the skin disease dataset ($B_{attack} = 32$, $A_{target} = 5\%$)

Model	Original Accuracy [%]	BFA				ZeBRA			
		Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Avg. Accuracy After Attack	Bit Flips (Best)	Bit Flips (Worst)	Mean / Stdev	Avg. Accuracy After Attack
ResNet-50	63.23	3	29	9.7 / 6.61	4.68	5	10	7.53 / 1.09	4.71
MobileNetV2	54.05	2	9	3.5 / 2.04	4.58	2	4	2.7 / 0.64	4.82

(22.4 bits) is 1.7× lower than the BFA (38.7 bits). Thus, the ZeBRA shows better attack performance even at the fortified DNNs.

Table 6: Top-1 test accuracy of fortified ResNet-20 on CIFAR-10

N_Q	$\lambda_{cluster}$ = 0	0.0001	0.0005	0.001
8bit	92.41	92.06	91.64	91.06
6bit	92.18	91.78	91.29	90.87
4bit	87.59	88.87	90.23	89.33

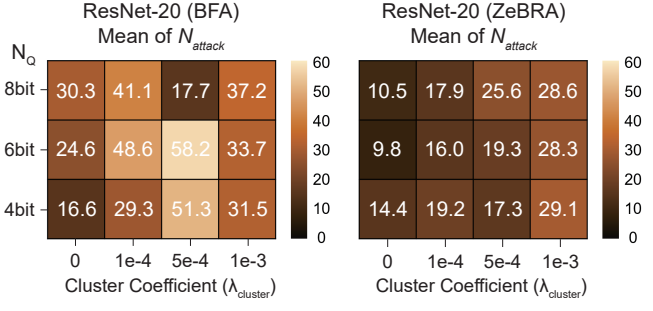


Figure 3: The average number of bit flips to destroy the fortified ResNet-20 on CIFAR-10 using the BFA or ZeBRA.

References

[1] DermNet NZ – all about the skin. URL https://dermnetnz.org/.

[2] Jun Cheng. Brain tumor dataset, April 2017. URL https://figshare.com/articles/brain_tumor_dataset/1512427/5.

[3] Zhezhi He, Adnan S. Rakin, Jingtao Li, Chaitali Chakrabarti, and Deliang Fan. Defending and harnessing the bit-flip based adversarial weight attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14083–14091, 2020. doi: 10.1109/CVPR42600.2020.01410.

[4] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[5] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. *arXiv:1801.04381*, 2019.

[6] Zar Nawab Khan Swati, Qinghua Zhao, Muhammad Kabir, Farman Ali, Zakir Ali, Saeed Ahmed, and Jianfeng Lu. Brain tumor classification for mr images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75:34–46, 2019.

[7] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019.

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008.