WHAT MAKES VISION TRANSFORMERS ROBUST TOWARDS BIT-FLIP ATTACK?

Xuan Zhou & Dake Chen & Peter Beerel
Ming Hsieh Department of Electrical and Computer Engineering
University of Southern California
Los Angeles CA 90089, USA
{zhouxuan,dakechen,pabeerel}@usc.edu

Souvik Kundu
Intel Labs
souvikk.kundu@intel.com

ABSTRACT

The bit-flip attack (BFA) is a well-studied assault that can dramatically degrade the accuracy of a machine learning model by flipping a small number of bits in the model parameters. Numerous studies have focused on enhancing the performance of BFA and mitigating their effects on traditional Convolutional Neural Networks (CNNs). However, there remains a lack of understanding regarding the security of vision transformers against BFA. In our work, we conduct various experiments on vision transformer models and discover that the flipped bits are concentrated in the MLP layers, specifically in the initial and final several blocks. Furthermore, we find an inverse relationship between the size of the transformer model and its robustness. Our findings in this study can aid in refining defense techniques, targeting them towards areas in vision transformer models that are particularly vulnerable to BFA.

1 INTRODUCTION

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) has garnered significant attention, primarily due to its innovative approach to image analysis that sets it apart from traditional neural networks. Unlike conventional convolutional neural networks (CNNs) (Waibel et al., 2013; Zhang et al., 1988; Krizhevsky et al., 2012) that rely on local feature extraction, ViT adopts the transformer architecture, which includes attention layers, originally designed for natural language processing, to process images as sequences of patches (Vaswani et al., 2017). This enables global context understanding and the capture of long-range spatial dependencies within images, leading to remarkable improvements in various vision tasks. The ability of ViTs to efficiently scale with increased data and compute resources further underscores their superiority, making them a pivotal development in the field.

Given the increasing prevalence of ViTs in critical applications, ranging from medical imaging (Shamshad et al., 2023; Chen et al., 2021b; Dalmaz et al., 2022) to autonomous driving (Ando et al., 2023; Prakash et al., 2021), their security and robustness have become paramount. As these models are integrated into more systems, the need to safeguard them against potential attacks is urgent. Current research indicates a growing number of sophisticated attacks targeting neural networks (Liu et al., 2018; Su et al., 2019; Zügner et al., 2018; Liu et al., 2020b). Among these attacks, the BFA (Rakin et al., 2019) is a particularly insidious threat to the integrity and security of neural networks, often overshadowing other forms of attacks such as adversarial and back-door attacks. Unlike adversarial attacks, which typically require input manipulation to deceive a neural network, BFA targets the physical memory of the hardware running the neural network. BFA executes these bit flips through the row hammer attack, a hardware fault injection technique targeting dynamic random-access memory (DRAM). By flipping a limited number of bits in the neural network's parameters stored in memory, an attacker can induce a catastrophic decrease in the network's prediction accuracy. This subtlety makes BFA especially dangerous, as they can be hard to detect.

Despite the significant advancements and widespread adoption of Vision Transformers (ViTs) in the field of computer vision, there remains a notable gap in the understanding of their vulnerability to BFA, as most previous works focus on BFA on traditional CNNs, leaving the vulnerability of vision transformer towards BFA unstudied. This work, to the best of our knowledge, is the first that analyzes the performance of BFA on vision transformers and explores the characteristics that can make vision transformer robust to BFA. We perform BFA on three sizes, tiny, small, and base, of two mainstream vision transformer models, standard ViT and DeiT. We then analyze the distribution of flipped bits in vision transformer models and compare the results of BFA on standard ViT and Data-efficient Image Transformer (DeiT) to find which characteristics can be attributed to their BFA vulnerability.

The rest of this paper is as follows: Section 2 gives an overview of prior works. Section 3 describes the experimental setup and presents ouir results and analysis. Finally, Section 4 concludes the paper.

2 Preliminaries

2.1 Vision transformer

The Vision Transformer represents a cutting-edge architecture that leverages the attention mechanism to efficiently extract important input features (Vaswani et al., 2017). The standard vision transformer consists of a patch embedding layer, 11 identical blocks, and a final MLP layer for classification. Each block contains attention and Multi-Layer Perceptron (MLP) layers. The layer organization is the same but varies in size, with configurations like Base, Small, and Tiny differing in the dimensionality of layers. Since its introduction in 2021 (Dosovitskiy et al., 2020), various adaptations of ViT have emerged, demonstrating its versatility and potential for innovation. Notable variants include the DeiT, which enhances model efficiency through token distillation (Touvron et al., 2021), the Swin Transformer, which optimizes representation through shifted windowing techniques (Liu et al., 2021), and the Convolutional vision Transformer (CvT) that incorporates convolutional operations into the transformer framework for improved performance (Wu et al., 2021). Their widespread adoption has underscored the importance of addressing security concerns to ensure the integrity and reliability of applications utilizing this model.

2.2 Bit-flip attack

A BFA (Rakin et al., 2019) aims to significantly lower the accuracy of a model to nearly random guesses by flipping a small number of bits in the model's parameters, identified by its progressive bit search algorithm. The core concept of the BFA involves using gradient ranking to pinpoint the bits that are likely to cause the most significant accuracy degradation. In addition, previous work has enhanced the BFA by improving the row hammer attack (Yao et al., 2020) and progressive bit search algorithm (Rakin et al., 2021) as well as modifying the BFA to enable the attack without access to training or testing data (Ghavami et al., 2022b).

While the original BFA was applied to CNNs, some works have expanded the BFA to other kinds of models, including pruned DNNs (Lee & Chandrakasan, 2022), transformers (Cai et al., 2021), and even DNN executables compiled by DL compilers (Chen et al., 2023). Besides BFA that aims to degrade accuracy, there is also a BFA variant, targeted BFA (Chen et al., 2021a; Rakin et al., 2020; 2022), that will not only reduce accuracy but also mislead the input to be classified into a preset class. This targeted BFA has been applied to multiple models, including binary pattern network (Roohi & Angizi, 2022) and transformers (Liu et al., 2023c).

Given the strong threat of BFA on neural networks, numerous strategies have been developed to counteract such attacks in traditional DNNs. Many of these approaches involve manipulating the weights to enhance the models' resistance to the attack. Techniques such as binary neural network (BNN) (Siraj Rakin et al., 2021; He et al., 2020), weight reconstruction (Li et al., 2020), smart bit flip (Ghavami et al., 2022a), random rotation of weight bits (Liu et al., 2023a), and quantization (Liu et al., 2023a; Stutz et al., 2023) have been demonstrated to be effective. Additionally, encoding the weights provides another form of protection (Javaheripi et al., 2022; Guo et al., 2021; Özdenizci & Legenstein, 2022; Li et al., 2021). Furthermore, hardware-based defense mechanisms are available to counteract row hammer attacks (Zhou et al., 2023; Gongye et al., 2023), which underpin BFA.

Notably, in targeted BFA, the flipped bits typically affect the final layers. Therefore, an efficient early exit strategy has been identified, enabling models to deliver classification outcomes in earlier stages of the model, thus thwarting targeted BFA (Wang et al., 2023). There is also innovative work introducing honey neurons into models to lure attackers by targeting these decoy neurons instead of the crucial functional ones (Liu et al., 2023b). Beyond minimizing the impact of BFA, significant efforts are also directed towards detecting such attacks (Yang et al., 2022; Liu et al., 2020a).

3 Analysis of Bit-Flip Susceptibility

3.1 Models, datasets, and attack metrics

To demonstrate BFA (Rakin et al., 2019) on ViT, we choose two mainstream ViT models, standard ViT (Vaswani et al., 2017) and DeiT (Touvron et al., 2021). For each ViT model, we perform experiments on three different sizes: tiny, small, and base. We take two visual datasets, CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2010), for object classification tasks. Both CIFAR10 and CIFAR100 contain 60K RGB images of size 32 × 32 × 3. There are 10 classes in CIFAR10 while 100 classes in CIFAR100. The valid flipped bits are those causing the accuracy to degrade before it reaches random guess, which is 11% and 1.1% separately for CIFAR10 and CIFAR100.

3.2 Bit-flip attack on models of different sizes

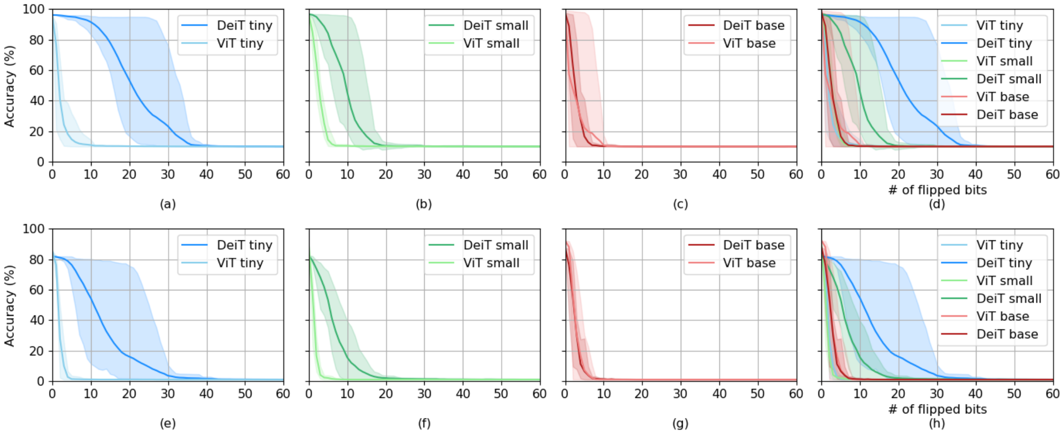


Figure 1: The top 1 accuracy of BFA after performing 0 to 60 bit flips. (abcd) CIFAR-10 (efgh) CIFAR-100.

Observation 1: Larger model size will lead to a more vulnerable model towards bit-flip attack.

Figure 1(d) and (e) reveal that DeiT Tiny, the most compact of the three models, consistently exhibits the highest resilience to BFA. Conversely, DeiT Base, the largest model, is the most susceptible to such attacks. DeiT Small occupies an intermediate position in both size and robustness against BFA. However, as Figure 1 indicates, ViT models are excessively vulnerable to BFA, leading to overlapping curves for the tiny, small, and base ViT models.

Observation 2: DeiT is more robust than ViT of the same model size. The smaller the model size is, the more robust DeiT is compared with ViT of the same size.

Figure 1(a-c, e-g) demonstrate that DeiT consistently outperforms ViT in terms of robustness when comparing models of identical size. The only distinction between DeiT and ViT is the presence of a distillation token in DeiT (Touvron et al., 2021; Dosovitskiy et al., 2020), which significantly enhances the resilience of DeiT towards BFA. This suggests incorporating additional distillation features into vision transformer models could enhance their defense against BFA. Furthermore, DeiT's robustness advantage over ViT increases as the model size decreases. Both the DeiT Base and ViT Base models are extremely vulnerable, and the performance gap between them is narrow, indicating a closer level of robustness.

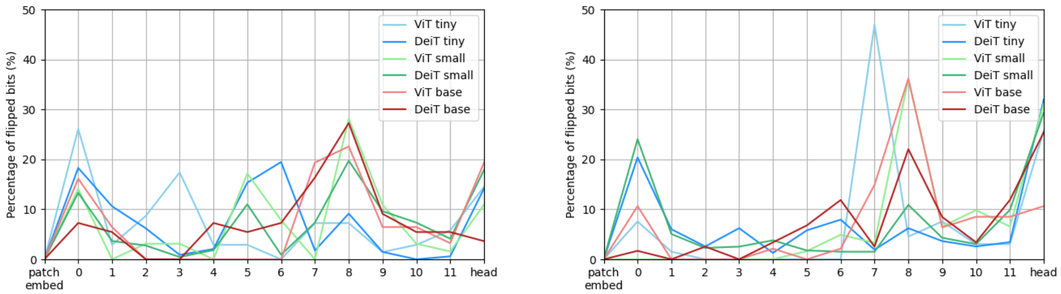
Table 1: Disctribution of flipped bits in types of layers

Dataset	Model	Flipped bits distributed in layers(%)		
		attention	MLP	head
CIFAR-10	ViT-Tiny	15.942	69.565	14.493
	ViT-Small	12.500	76.563	10.938
	ViT-Base	22.581	58.065	19.355
	DeiT-Tiny	23.304	62.537	14.159
	DeiT-Small	9.174	72.936	17.890
	DeiT-Base	12.727	83.636	3.636
CIFAR-100	ViT-Tiny	8.333	91.667	0
	ViT-Small	5.263	94.737	0
	ViT-Base	16.129	80.645	3.226
	DeiT-Tiny	41.860	52.907	5.233
	DeiT-Small	34.653	65.347	0
	DeiT-Base	2.857	97.143	0

3.3 Distribution of flipped bits

Observation 3: Flipped bits concentrate in MLP layers.

The data presented in Table 1 from our experiments indicate a tendency for bit flips to concentrate within MLP layers. Notice that no flipped bit falls in the patch embed layer in any of the experiments. The attention layers, which consist of a Query, Key, Value (QKV) framework, which collaborate to dynamically distribute attention across various segments of the input data, contain certain redundancies. Due to redundancy within the QKV structure, a bit flip in attention layers generally results in less reduction in accuracy than a bit flip in MLP layers.



(a)

(b)

Figure 2: Distribution of flipped bits in blocks on dataset (a) CIFAR-10 (b) CIFAR-100.

Observation 4: Flipped bits concentrate on the layers in the first block and the last few blocks.

Errors in the initial and final layers of the model can significantly harm its performance. This is because errors at the initial layers can accumulate throughout the model's computations, and errors in the final layers can more directly affect classification outcomes. Interestingly, previous analysis of BFA on DNNs also found front-end layers to be highly vulnerable (He et al., 2020), but our observations suggest the vulnerability of later layers appears to be more pronounced in vision transformers.

Past methods for defending against BFA often resulted in increased power (He et al., 2020; Li et al., 2020; Liu et al., 2020a) or lowered baseline classification accuracy (He et al., 2020; Siraj Rakin et al., 2021). However, with new insights into the distribution of flipped bits, these defense strategies can be enhanced by focusing protection on the most susceptible layers—specifically, the MLP layers in the first block and the last few blocks. This approach thus promises to reduce the cost of these defenses.

4 Conclusions

This paper studies the factors contributing to the resilience of vision transformer models against BFA, paving the way for the development of more secure vision transformers. Through experimental analysis, it has been identified that the robustness of vision transformers to BFA is affected by several factors, notably the size of the model and the application of distillation techniques. Furthermore, the investigation reveals a distinct pattern in the distribution of flipped bits, with a propensity for these bits to cluster within the initial and final layers of the architecture. This insight could be instrumental in formulating low-cost defensive strategies against BFA.

REFERENCES

Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5240–5250, 2023.

Kunbei Cai, Md Hafizul Islam Chowdhuryy, Zhenkai Zhang, and Fan Yao. Seeds of seed: Nmt-stroke: Diverting neural machine translation through hardware-based faults. In *2021 International Symposium on Secure and Private Execution Environment Design (SEED)*, pp. 76–82, 2021. doi: 10.1109/SEED51797.2021.00019.

Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Proflip: Targeted trojan attack with progressive bit flips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7718–7727, October 2021a.

Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021b.

Yanzuo Chen, Zhibo Liu, Yuanyuan Yuan, Sihang Hu, Tianxiang Li, and Shuai Wang. Unveiling signle-bit-flip attacks on dnn executables. *arXiv preprint arXiv:2309.06223*, 2023.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Behnam Ghavami, Seyd Movi, Zhenman Fang, and Lesley Shannon. Stealthy attack on algorithmic-protected dnns via smart bit flipping. In *2022 23rd International Symposium on Quality Electronic Design (ISQED)*, pp. 1–7, 2022a. doi: 10.1109/ISQED54688.2022.9806152.

Behnam Ghavami, Mani Sadati, Mohammad Shahidzadeh, Zhenman Fang, and Lesley Shannon. Bdfa: A blind data adversarial bit-flip attack on deep neural networks, 2022b.

Cheng Gongye, Yukui Luo, Xiaolin Xu, and Yunsi Fei. Hammerdodger: A lightweight defense framework against rowhammer attack on dnns. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2023. doi: 10.1109/DAC56929.2023.10247671.

Yanan Guo, Liang Liu, Yueqiang Cheng, Youtao Zhang, and Jun Yang. Modelshield: A generic and portable framework extension for defending bit-flip based adversarial weight attacks. In *2021 IEEE 39th International Conference on Computer Design (ICCD)*, pp. 559–562, 2021. doi: 10.1109/ICCD53106.2021.00090.

Zhezhi He, Adnan Siraj Rakin, Jingtao Li, Chaitali Chakrabarti, and Deliang Fan. Defending and harnessing the bit-flip based adversarial weight attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Mojan Javaheripi, Jung-Woo Chang, and Farinaz Koushanfar. Acchashtag: Accelerated hashing for detecting fault-injection attacks on embedded neural networks. *ACM Journal on Emerging Technologies in Computing Systems*, 19(1):1–20, 2022.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 (canadian institute for advanced research). *URL http://www.cs.toronto.edu/ kriz/cifar.html*, 5(4):1, 2010.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Kyungmi Lee and Anantha P. Chandrakasan. Sparsebfa: Attacking sparse deep neural networks with the worst-case bit flips on coordinates. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4208–4212, 2022. doi: 10.1109/ICASSP43922.2022.9747337.

Jingtao Li, Adnan Siraj Rakin, Yan Xiong, Liangliang Chang, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. Defending bit-flip attack through dnn weight reconstruction. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2020. doi: 10.1109/DAC18072.2020.9218665.

Jingtao Li, Adnan Siraj Rakin, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. Radar: Run-time adversarial weight attack detection and accuracy recovery. In *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 790–795, 2021. doi: 10.23919/DATE51398.2021.9474113.

Liang Liu, Yanan Guo, Yueqiang Cheng, Youtao Zhang, and Jun Yang. Generating robust dnn with resistance to bit-flip based adversarial weight attack. *IEEE Transactions on Computers*, 72(2): 401–413, 2023a. doi: 10.1109/TC.2022.3211411.

Qi Liu, Wujie Wen, and Yanzhi Wang. Concurrent weight encoding-based detection for bit-flip attack on neural network accelerators. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2020*, 2020a.

Qi Liu, Jieming Yin, Wujie Wen, Chengmo Yang, and Shi Sha. Neuropots: Realtime proactive defense against bit-flip attacks in neural networks. 2023b.

Yepeng Liu, Bo Feng, and Qian Lou. Trojtext: Test-time invisible textual trojan insertion. *arXiv preprint arXiv:2303.02242*, 2023c.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 182–199. Springer, 2020b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Ozan Özdenizci and Robert Legenstein. Improving robustness against stealthy weight bit-flip attacks by output code matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13388–13397, June 2022.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Bit-flip attack: Crushing neural network with progressive bit search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13198–13207, 2020.

Adnan Siraj Rakin, Yukui Luo, Xiaolin Xu, and Deliang Fan. Deep-Dup: An adversarial weight duplication attack framework to crush deep neural network in Multi-Tenant FPGA. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1919–1936. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL `https://www.usenix.org/conference/usenixsecurity21/presentation/rakin`.

Adnan Siraj Rakin, Zhezhi He, Jingtao Li, Fan Yao, Chaitali Chakrabarti, and Deliang Fan. T-BFA: Targeted bit-flip adversarial weight attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7928–7939, 2022. doi: 10.1109/TPAMI.2021.3112932.

Arman Roohi and Shaahin Angizi. Efficient targeted bit-flip attack against the local binary pattern network. In *2022 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pp. 89–92, 2022. doi: 10.1109/HOST54066.2022.9839959.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, pp. 102802, 2023.

Adnan Siraj Rakin, Li Yang, Jingtao Li, Fan Yao, Chaitali Chakrabarti, Yu Cao, Jae-sun Seo, and Deliang Fan. RA-BNN: Constructing robust & accurate binary neural network to simultaneously defend adversarial bit-flip attack and improve accuracy. *arXiv e-prints*, pp. arXiv–2103, 2021.

David Stutz, Nandhini Chandramoorthy, Matthias Hein, and Bernt Schiele. Random and adversarial bit error robustness: Energy-efficient and secure dnn accelerators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3632–3647, 2023. doi: 10.1109/TPAMI.2022.3181972.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. In *Backpropagation*, pp. 35–61. Psychology Press, 2013.

Jialai Wang, Ziyuan Zhang, Meiqi Wang, Han Qiu, Tianwei Zhang, Qi Li, Zongpeng Li, Tao Wei, and Chao Zhang. Aegis: Mitigating targeted bit-flip attacks against deep neural networks. *arXiv preprint arXiv:2302.13520*, 2023.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, October 2021.

Li-Hsing Yang, Shin-Shan Huang, Tsai-Ling Cheng, Yi-Ching Kuo, and Jian-Jhih Kuo. Socially-aware collaborative defense system against bit-flip attack in social internet of things and its online assignment optimization. In *2022 International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–10, 2022. doi: 10.1109/ICCCN54977.2022.9868899.

Fan Yao, Adnan Siraj Rakin, and Deliang Fan. DeepHammer: Depleting the intelligence of deep neural networks through targeted chain of bit flips. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1463–1480. USENIX Association, August 2020. ISBN 978-1-939133-17-5. URL `https://www.usenix.org/conference/usenixsecurity20/presentation/yao`.

Wei Zhang, Jun Tanida, Kazuyoshi Itoh, and Yoshiki Ichioka. Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of annual conference of the Japan Society of Applied Physics*, volume 564. Montreal, CA, 1988.

Ranyang Zhou, Sabbir Ahmed, Adnan Siraj Rakin, and Shaahin Angizi. Dnn-defender: An in-dram deep neural network defense mechanism for adversarial weight attack, 2023.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2847–2856, 2018.