Towards Robust RRAM-based Vision Transformer Models with Noise-aware Knowledge Distillation

Wenyong Zhou[1], Zhengwu Liu[1,*], Taiqiang Wu[1], Chenchen Ding[1], Yuan Ren[1], Ngai Wong[1,*]
[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong
[*]Corresponding authors. Email: {zwliu, nwong}@eee.hku.hk

***Abstract*—Resistive random-access memory (RRAM)-based compute-in-memory (CIM) systems show promise in accelerating Transformer-based vision models but face challenges from inherent device non-idealities. In this work, we systematically investigate the vulnerability of Transformer-based vision models to RRAM-induced perturbations. Our analysis reveals that earlier Transformer layers are more vulnerable than later ones, and feed-forward networks (FFNs) are more susceptible to noise than multi-head self-attention (MHSA). Based on these observations, we propose a noise-aware knowledge distillation framework that enhances model robustness by aligning both intermediate features and final outputs between weight-perturbed and noise-free models. Experimental results demonstrate that our method improves accuracy by up to 1.54% and 1.49% on ViT and DeiT models under various noise conditions compared to their vanilla counterparts.**

***Index Terms*—Transformer-based Vision Models, RRAM, Compute-in-Memory, Knowledge Distillation**

I. Introduction

Transformer-based vision models have transformed computer vision by adapting self-attention mechanisms from natural language processing [1], [2]. These models process image patches as tokens and use multi-head self-attention (MHSA) instead of traditional convolutions, enabling superior performance across various tasks through better modeling of long-range relationships [3]. However, their computational intensity challenges conventional computing architectures. RRAM-based compute-in-memory (CIM) systems emerge as a promising solution, offering high density and energy efficiency through parallel matrix-vector multiplications [4]. Despite these advantages, performance degradation, as shown in Fig. 1 due to RRAM's inherent non-idealities remains a significant challenge. While previous work addresses RRAM-based CNN robustness, improving Transformer models presents unique challenges due to their distinct architecture [5].

To address this issue, we conduct comprehensive experiments on the vulnerability of ViT [1] and DeiT [2] models and propose a noise-aware knowledge distillation (KD) framework to enhance their robustness. Our analysis reveals that earlier Transformer layers are more vulnerable than later ones, and feed-forward networks (FFNs) exhibit higher noise susceptibility than MHSA. Based on these findings, our noise-aware KD framework aligns both intermediate features and final outputs between weight-perturbed and noise-free models. Experimental results demonstrate that our method achieves accuracy improvements of up to 1.54% and 1.49% on ViT and DeiT models compared to their vanilla counterparts.

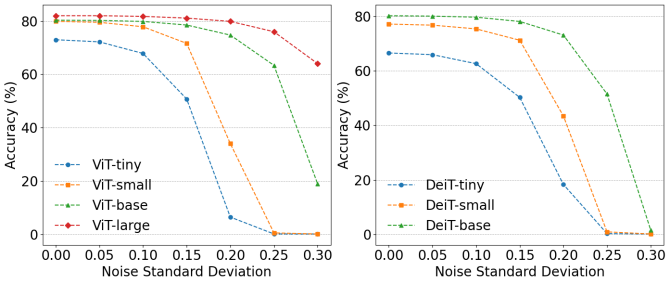

Fig. 1. Model accuracy degradation of ViT and DeiT under RRAM-induced weight perturbations.

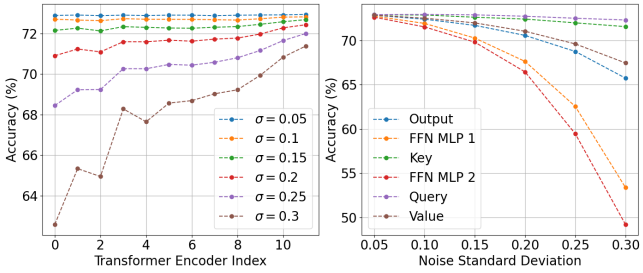

Fig. 2. Layer-wise vulnerability analysis of ViT-tiny showing accuracy degradation under different noise levels across encoder positions and component types.

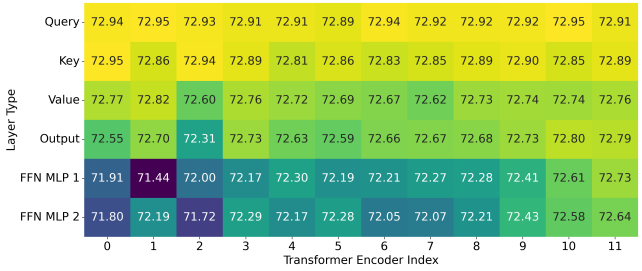

Fig. 3. Accuracy heatmap of ViT-tiny under weight perturbations across different model components.

II. Methodology

For fair comparison, we use pre-trained models from HuggingFace and model noise using a lognormal distribution, following common practices [6]. We analyze model sensitivity by perturbing weights in both MHSA and FFN linear layers across different dimensions, revealing the following characteristics.

978-3-9826741-0-0/DATE25/© 2025 EDAA

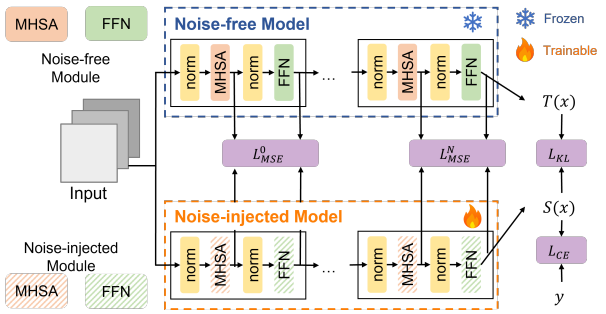

Fig. 4. Our proposed noise-aware KD framework for Transformer-based vision models.

First, as shown in Fig. 1, model size significantly impacts robustness, with ViT and DeiT base/large variants substantially outperforming their tiny/small counterparts, especially under high noise levels. At similar model sizes, DeiT models show lower performance under both slight and high noise but perform better under moderate noise conditions. Second, as shown in Figs. 2 and 3, the position of Transformer encoders demonstrates asymmetric impact on final accuracy, with shallow encoder layers showing greater accuracy degradation under the same noise level compared to deeper layers. This can be attributed to the error amplification effect in deep neural networks. Third, as shown in Figs. 2 and 3, linear layers in different modules exhibit varying degrees of robustness: FFN layers demonstrate greater sensitivity than MHSA layers, while within MHSA, query and key layers show higher robustness compared to value and output layers.

Based on these observations, we propose a noise-aware KD framework to enhance model robustness [5]. Rather than training models from scratch, which is computationally expensive, our framework utilizes pre-trained models and finetunes their noisy variants for only a few epochs. In this framework, we designate the noise-free model as the teacher and its noise-perturbed counterpart as the student, aiming to transfer knowledge from teacher to student during finetuning. Since both models share identical architectures and differ only in that the student model's weights are perturbed by noise during each forward pass, we can effectively align both output and intermediate features. This alignment allows the student model to mitigate variations at each layer rather than allowing them to accumulate in the final output, as illustrated in Fig. 4. The total loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta\mathcal{L}_{KL} + \gamma\mathcal{L}_{MSE} \quad (1)$$

where β and γ are weights balancing the three loss terms, and $\mathcal{L}_{CE}$, $\mathcal{L}_{KL}$, and $\mathcal{L}_{MSE}$ represent the cross-entropy loss for the original task, KL divergence loss for output alignment, and mean squared error for intermediate feature alignment, respectively.

III. Experiments

Due to limited prior work on vision Transformer robustness, we compare our method against vanilla ViT and DeiT models. We focus on their tiny and small variants, which demonstrated higher vulnerability to noise in our previous experiments.

TABLE I
Accuracy Comparison of ViT and DeiT Models Under Different Noise Levels

Model	Method	Device Variation		
		$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.15$
ViT-tiny	Baseline	72.18%	67.83%	50.75%
	Ours	**72.94%**	**68.72%**	**52.29%**
ViT-small	Baseline	79.51%	77.84%	71.61%
	Ours	**79.94%**	**78.31%**	**72.49%**
DeiT-tiny	Baseline	65.91%	62.62%	50.24%
	Ours	**66.68%**	**63.49%**	**51.73%**
DeiT-small	Baseline	76.74%	75.37%	71.21%
	Ours	**77.05%**	**75.89%**	**72.03%**

Table I presents the comparative results. Our method achieves consistent improvements across all models and noise conditions compared to vanilla baselines. The performance gains are more pronounced in smaller models, highlighting the method's potential for lightweight edge deployments. For instance, at $\sigma = 0.05$, the accuracy improvement increases from 0.43% to 0.76% when moving from ViT-small to ViT-tiny. Moreover, our method demonstrates better accuracy retention under higher noise levels. For example, the accuracy gain on DeiT-small increases from 0.31% to 1.22% as noise levels increase from $\sigma = 0.05$ to $\sigma = 0.15$.

IV. Conclusion

This work presents the first systematic investigation of RRAM-induced non-idealities on vision Transformers, revealing heightened vulnerability in early encoder layers and FFN modules. Based on these findings, we propose a noise-aware knowledge distillation framework that aligns both intermediate features and final outputs between weight-perturbed models and noise-free references. Our method achieves accuracy improvements of up to 1.54% and 1.49% for ViT and DeiT models, respectively, under various noise conditions.

Acknowledgement

This work was supported in part by the Theme-based Research Scheme (TRS) project T45-701/22-R, National Natural Science Foundation of China (62404187) and the General Research Fund (GRF) Project 17203224, of the Research Grants Council (RGC), Hong Kong SAR.

References

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.
[2] H. Touvron *et al.*, "Training Data-Efficient Image Transformers Distillation through Attention," in *ICML*, 2021.
[3] A. Vaswani *et al.*, "Attention is All You Need," in *NeurIPS*, 2017.
[4] Q. Zheng *et al.*, "Accelerating Sparse Attention with a Reconfigurable Non-volatile Processing-In-Memory Architecture," in *DAC*, 2023.
[5] G. Charan *et al.*, "Accurate Inference with Inaccurate RRAM Devices: Statistical Data, Model Transfer, and On-line Adaptation," in *DAC*, 2020.
[6] G. Jung *et al.*, "Cost- and Dataset-free Stuck-at Fault Mitigation for ReRAM-based Deep Learning Accelerators," in *DATE*, 2021.