Adversarial Robustness of Vision Transformers

Karan Gandhi
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110157@iitgn.ac.in

Anurag Signh
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110035@iitgn.ac.in

Arjun Dikshit
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110040@iitgn.ac.in

Aarsh Wankar
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110003@iitgn.ac.in

Abstract

Vision Transformers (ViTs) have demonstrated strong performance in image recognition but remain highly vulnerable to frequency-based adversarial attacks. This project systematically analyzes how adversarial perturbations impact ViTs, particularly through disruptions in positional encodings. We investigate the role of different encoding strategies in this vulnerability and propose defenses to enhance robustness. By examining attention maps and token representations, we aim to gain deeper insights into the adversarial weaknesses of ViTs and contribute to the development of more resilient architectures.

Keywords

Vision Transformers, Adversarial Attacks, Spectral Attacks, Positional Encodings

1 Introduction

Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved remarkable success in image recognition tasks [6, 7, 13]. However, they are highly susceptible to adversarial attacks [5], such as small, carefully crafted perturbations to the image at the testing time that lead the model into incorrect classifications. More recently, Vision Transformers (ViTs) [4] have gained popularity due to their impressive performance in image recognition. These ViTs have shown greater resilience to spatial perturbation attacks compared to CNNs [1]. However, they remain highly vulnerable to frequency-based attacks [8]. In this project, we aim to systematically investigate this vulnerability by analyzing and visualizing how adversarial perturbations impact key components of ViTs, particularly their attention scores. We also hypothesize that ViTs are susceptible to frequency-based attacks because of the addition of sinusoidal positional encodings, which are disrupted by spectral noise, leading to degraded performance. To the best of our knowledge, this vulnerability has not been explored.

Thus, we aim to validate this hypothesis by analyzing how different adversarial attacks affect ViTs with different positional encodings. This study aims to provide deeper insights into the robustness of ViTs and contribute to the development of more resilient architectures against adversarial threats.

2 Related Work

Traditional adversarial attacks have focused on spatial perturbations, modifying pixel values directly. However, recent research has highlighted spectral-domain attacks, which target the Fourier-transformed representation of images.

A unified adversarial attack framework has been introduced to apply perturbations selectively in both spatial and spectral domains, enabling frequency-selective adversarial attacks on ViTs. Unlike CNNs, which struggle against high-frequency noise, ViTs are more sensitive to perturbations in the low and intermediate-frequency bands, making them highly susceptible to frequency-based adversarial attacks. [8]

Frequency-domain attacks provide new insights into ViT vulnerabilities as mentioned in [6], showing that the Phase perturbations are significantly more damaging than magnitude perturbations for ViTs. This is because ViTs heavily depend on phase information for classification, while CNNs rely more on magnitude information. Also, it mentions that the effectiveness of Attacks depends on model size and training data. Thus, larger models (ViT-L) show slightly improved robustness over smaller models (ViT-B), but remain highly vulnerable to low-frequency phase attacks. It concludes that the robustness advantage of ViTs over CNNs is attack-dependent. When considering standard pixel-based perturbations (FGSM, PGD), ViTs perform better, but under spectral attacks, they perform worse than CNNs. [8]

Adversarial pretraining has been shown to be effective in preventing adversarial attacks and also improve generalization ability during classification. There are two main categories: Memory-free methods with instance-wise perturbations and Memory-based methods using feature-level adversaries.

[11].

3 Methodology

In this study, we aim to examine the vulnerability of Vision Transformers (ViTs) to adversarial perturbations. First, we will conduct adversarial experiments on both CNNs and ViTs to analyze how adversarial noise affects models in both the frequency and spatial domains. We will begin by replicating existing findings that show CNNs are more vulnerable to both types of perturbations, while ViTs exhibit greater susceptibility to frequency-based attacks. Once these baseline results are established, we will investigate the underlying reason for ViTs' sensitivity to frequency-based attacks. To do this, we will analyze token representations after applying positional encodings and examine attention maps to understand how these attacks affect ViTs' internal mechanisms.

Next, we will test our hypothesis by evaluating the performance of ViTs with different positional encoding strategies, including Sinusoidal Positional Embeddings, Coordinate Positional Embeddings, Rotary Positional Embeddings (RoPE), and Conditional Positional Embeddings (CPE). A potential challenge here is isolating the effect of positional encodings from other architectural components that might also contribute to adversarial vulnerability. Furthermore, the choice of evaluation metrics will be crucial to ensure a fair comparison between different positional encodings under various attack scenarios. To further interpret our findings, we will conduct additional experiments, such as analyzing the change in the change pairwise L2 norm of tokens after positional embeddings in the perturbed image, to better understand the impact of frequency-based adversarial noise.

Finally, we will explore defense strategies to enhance ViTs' robustness against adversarial attacks. This includes investigating alternative positional encoding techniques like RoFormer and evaluating how different embeddings impact model resilience against spectral perturbations. Additionally, we will assess the effectiveness of adversarial pretraining methods, including MoCo, SimCLR, and AdCo, by subjecting ViTs to these techniques before applying adversarial attacks.

Through these experiments, we aim to gain deeper insights into ViTs' adversarial vulnerabilities and propose effective mitigation strategies to improve their robustness against spectral perturbations. By addressing these challenges, we hope to provide a clearer understanding of the role of positional encodings in adversarial susceptibility and contribute to the development of more resilient vision transformer architectures.

4 Datasets and Compute Requirements

The datasets that we would use is:

- Normal Image dataset
  - ImageNet [3]
  - CIFAR - 10 [9]
  - CIFAR - 100 [10]
- Adversarially perturbed dataset
  - Adversarial Dataset [12]
  - DAmageNet [2]

Given the computational complexity of training and evaluating Vision Transformers under adversarial settings, we require access to high-performance compute resources, including GPUs/TPUs, to efficiently conduct our experiments. Adequate compute power will be crucial for processing large-scale datasets, generating adversarial examples, and analyzing the impact of adversarial perturbations on ViTs.

We would greatly appreciate access to suitable computational resources to ensure the successful execution of our project.

5 Project Timeline

(1) **Week 1: Literature Review & Baseline Implementation** Conduct a literature review on adversarial attacks, ViTs, and positional encodings. Implement FGSM, PGD, and C&W attacks on CNNs and ViTs. Train models on ImageNet, CIFAR-10, and CIFAR-100. Validate CNNs' vulnerability to spatial perturbations and ViTs' susceptibility to spectral attacks.

(2) **Week 2: Frequency-Based Adversarial Attacks** Implement spectral-domain attacks targeting ViTs. Compare the impact of spatial and frequency-based perturbations. Visualize attention maps to analyze how attacks disrupt different model components. Investigate the effect of adversarial perturbations on positional encodings.

(3) **Week 3: Effect of Positional Encodings** Implement Sinusoidal, RoPE, and CPE positional encodings. Evaluate their impact on model robustness under adversarial attacks. Analyze variations in L2 norm of token embeddings before and after encoding. Study how encoding disruptions correlate with accuracy degradation.

(4) **Week 4: Defense Mechanisms** Implement adversarial training and inverse perturbation techniques. Assess their effectiveness in mitigating attacks. Compare improvements in robustness across different defense strategies. Conduct an ablation study to determine optimal combinations of encoding and defense methods.

(5) **Week 5: Performance Evaluation & Visualization** Analyze ViTs' performance under different attack scenarios. Visualize findings using attention heatmaps and frequency-response plots. Compare robustness of ViTs and CNNs. Summarize key insights on vulnerabilities and potential countermeasures.

(6) **Week 6: Final Report & Documentation** Compile all findings, results, and visualizations into a structured report. Summarize key takeaways and future research directions. Prepare a final presentation highlighting contributions. Finalize documentation and submit the project report.

6 Conclusion

Through our study, we aim to provide a deeper understanding of how adversarial perturbations impact ViTs at a structural level. By systematically analyzing the influence of frequency-based attacks on different positional encoding methods, we expect to highlight the key factors contributing to ViT vulnerabilities. Furthermore, our investigation into defense strategies will aid in designing more robust ViT architectures. The insights gained from this research will not only enhance adversarial robustness but also contribute to the broader field of secure deep learning. Given the computational demands of this research, access to high-performance GPUs/TPUs is essential for efficient training and evaluation. This work will serve as a foundation for future studies aiming to improve ViT security and generalization in real-world applications.

References

[1] Ahmed Aldahdooh, Wassim Hamidouche, and Olivier Deforges. 2021. Reveal of vision transformers robustness against adversarial attacks. *arXiv preprint arXiv:2106.03734* (2021).

[2] Sizhe Chen, Jiawei Gu, Tianyu He, Rui Qian, Yinpeng Chen, and Cihang Xie. 2023. DAmageNet: Attacking Object Detection in the Physical World. https://github.com/Sizhe-Chen/DAmageNet

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848

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

[5] Diego Gragnaniello, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. 2018. Analysis of adversarial attacks against CNN-based image forgery detectors. In *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 967–971.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[8] Gihyun Kim, Juyeop Kim, and Jong-Seok Lee. 2024. Exploring adversarial robustness of vision transformers in the spectral perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3976–3985.

[9] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. *Technical Report, University of Toronto* (2009). https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[10] Alex Krizhevsky and Geoffrey Hinton. 2009. CIFAR-100 Dataset. *Technical Report, University of Toronto* (2009). https://www.cs.toronto.edu/~kriz/cifar.html

[11] Guo-Jun Qi and Mubarak Shah. 2022. Adversarial pretraining of self-supervised deep networks: Past, present and future. *arXiv preprint arXiv:2210.13463* (2022).

[12] Bibhuti Bhusan Singh. 2022. Adversarial Dataset. https://www.kaggle.com/datasets/bibhutibhusansingh/adversarial-dataset

[13] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.