A study of RayS Attack on Vision Transformer Robustness using CIFAR10 Dataset

Venkata Abhiram Chitty
School of Data Science and Analytics
Kennesaw State University
Kennesaw, GA 30144
abhiramchitty@outlook.com

Abstract

Adversarial attacks significantly undermine the reliability of deep learning models in high-stakes applications, exposing their susceptibility to carefully crafted perturbations. This study investigates the robustness of a pre-trained Vision Transformer (ViT) model against the RayS attack, a query-efficient, decision-based black-box adversarial attack. Using the CIFAR-10 dataset, we evaluate the model's performance under varying perturbation constraints (epsMax = 0.031 and 0.062) and analyze its vulnerability to adversarial perturbations. The experimental results demonstrate a sharp decline in classification accuracy when subjected to adversarial examples, emphasizing the sensitivity of ViTs to imperceptible modifications. It visually compares clean and adversarial images and quantifies the attack's effectiveness in degrading model robustness.

The code for this study can be found in this GitHub [repo]

1 Introduction

Even with their remarkable performance on a wide range of computer vision tasks, deep learning models are nevertheless susceptible to adversarial attacks and well-crafted perturbations intended to skew model predictions without being noticeable to humans. When implementing these models in security-critical systems, this vulnerability presents serious difficulties.

Vision Transformers (ViTs) (1; 2) have emerged as powerful alternatives to Convolutional Neural Networks (CNNs) for image classification tasks. ViTs adapt the transformer architecture initially designed for natural language processing to handle visual data by splitting images into patches and processing them as token sequences. While ViTs have demonstrated state-of-the-art (SOTA) performance on various benchmarks, their robustness against adversarial attacks remains insufficiently explored. This work investigates the robustness of a pre-trained ViT model against the RayS attack, a query-efficient, decision-based black-box adversarial attack. The RayS attack is exciting because it requires minimal information about the target model, making it applicable in real-world scenarios where attackers have limited access to model details.

2 Background

2.1 Vision Transformer

Vision Transformer (1) has emerged as a strong alternative to CNNs in computer vision, offering superior computational efficiency and accuracy. Initially introduced in the 2017 paper "Attention is All You Need" (3) for NLP, ViT was adapted for vision tasks in "An Image is Worth 16x16 Words" (2). ViT processes images by dividing them into fixed-size patches, embedding them as tokens, adding

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

positional encodings, and passing them through a Transformer encoder with multi-head self-attention and multi-layer perceptrons (MLP). A classification head then produces the final output. ViT differs from CNNs by capturing global representations from shallow layers while maintaining local details. It leverages self-attention mechanisms rather than convolutional filters, making skip connections more impactful and preserving spatial information more effectively. ViTs perform exceptionally well on large datasets but require more data for optimal accuracy, whereas CNNs may be more effective for small datasets. ViTs also train faster at scale, making them efficient for high-volume applications.

ViTs have been applied in diverse fields, including image classification, segmentation, anomaly detection, action recognition, and autonomous driving. Their advantages include global context awareness, scalability, effective transfer learning, and interpretability through attention maps. In contrast to CNNs, they also have drawbacks such as high data requirements, high processing demands, and complicated deployment.

2.2 RayS Attack

RayS (4) is a hard-label adversarial attack method designed to efficiently and effectively generate adversarial examples against deep neural networks, even when the attacker has limited access to the model's outputs. Unlike traditional white-box or black-box attacks, which require gradient information or soft-label outputs, hard-label attacks only rely on the final predicted class of the model. This makes them more challenging but also more practical for real-world scenarios. The Ray Searching attack (RayS) addresses the inefficiency of previous hard-label attacks by reformulating the adversarial search problem from a continuous optimization task into a discrete search problem.

Instead of estimating gradients through zeroth-order methods, which can be computationally expensive and query-inefficient, RayS searches for adversarial perturbations by directly probing the decision boundary in a structured manner. It accomplishes this goal by examining a limited range of ray directions and using a fast-check technique to weed out pointless searches. It drastically lowers the number of queries needed to locate an adversarial example. Detecting "falsely robust" models—which seem safe from traditional adversarial attacks but are weak—is another helpful application of RayS. Several cutting-edge defence models have been seen to have their robust accuracy considerably reduced by it, exposing their actual vulnerability to hostile perturbations. Because of this, RayS is a powerful attack technique and a valuable instrument for assessing the steadfast resilience of machine learning models.

3 Experiments

3.1 Experimental Setup

The experiments were conducted using PyTorch (5) on the CIFAR-10 dataset (6). The pre-trained Vision Transformer model was loaded from the ViTRobust GitHub repository (7; 8), a reproduction of the ICCV 2021 paper On the Robustness of Vision Transformers to Adversarial Examples (9). The clean accuracy of the ViT model was evaluated before applying adversarial attacks.

3.2 Dataset

The CIFAR-10 dataset consists of 60,000 images across 10 classes. The dataset was preprocessed by normalizing images using the standard mean and standard deviation values for CIFAR-10. Images were resized and formatted to be compatible with the ViT model's input dimension (224 x 224 x 3). A subset of 100 clean, class-balanced images was selected from the validation set for adversarial evaluation, ensuring fair representation across all classes.

3.3 Model Implementation

The Vision Transformer model used in this study was pre-trained on ImageNet-21K (10) and fine-tuned on the CIFAR-10 dataset. The model was loaded using the TransformersModels.py script to ensure alignment with dataset requirements. The architecture, including multi-head self-attention and positional embeddings, was inspected for compatibility. The model's clean accuracy on the CIFAR-10 validation dataset was first evaluated to establish a baseline metric before introducing adversarial perturbations.

3.4 Adversarial Attack Implementation

The robustness of the ViT model was evaluated using the RayS attack, an efficient black-box adversarial attack method designed to identify minimal adversarial perturbations with limited access to the model. Unlike gradient-based white-box attacks, RayS only requires hard-label outputs from the model, making it highly effective in real-world black-box scenarios where internal model parameters are inaccessible. The attack performs a binary search along different perturbation directions to determine the closest decision boundary. RayS was executed using the AttackWrappersRayS.py script, and the attack was conducted under two different perturbation constraints to analyze the model's robustness under varying adversarial strengths:

1. **epsMax = 0.031:** This moderate perturbation level introduces subtle changes to the input images while aiming to induce misclassification. It represents a lower-bound adversarial scenario where the model is tested against minimal, nearly imperceptible alterations.
2. **epsMax = 0.062:** This higher perturbation level applies more pronounced adversarial modifications to the images, increasing the likelihood of successful misclassification while assessing the model's resilience against more potent attacks.

The attack was performed with a query limit of 3000 per image, ensuring a balance between computational efficiency and attack effectiveness. The following command was used to execute the RayS attack on a ViTL_16 model with a perturbation limit of 0.031 on 100 CIFAR-10 images. Adversarial images generated under these conditions were saved for further evaluation and visualization. The performance degradation of the model under adversarial attacks was measured to understand its vulnerability to black-box perturbations.

Robustness: Robustness specifically measures how well the model can withstand this black-box adversarial attack, which works by searching for minimal perturbations along random ray directions to cause misclassification.

4 Results

Table 1 illustrates the results of running a RayS attack on the Vit-L_16 model using 3000 queries and epsmax values of 0.031 and 0.062. We observe that the robust accuracy under the RayS attack is higher for lesser values of epsmax. Figures 1 to 10 illustrate the clean and adversarial images for epsmax value of 0.031.

Model	Eps Max	Queries Used	Robust Accuracy	Clean Accuracy
ViT-L_16	0.031	3000	0.47	0.991
ViT-L_16	0.062	3000	0.26	0.991

References

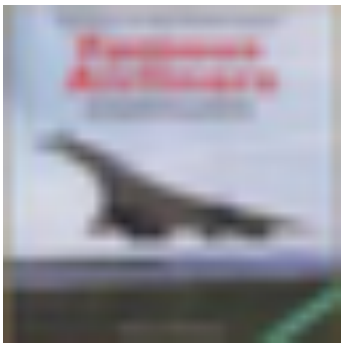
[1] Viso.ai, "Vision transformer (vit) – deep learning for computer vision." Online article, 2024. `https://viso.ai/deep-learning/vision-transformer-vit/`.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

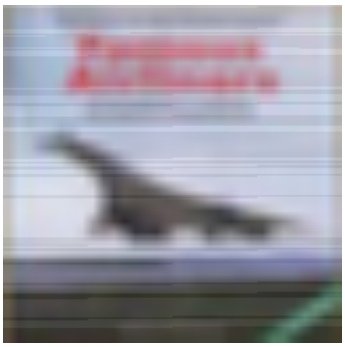
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] J. Chen and Q. Gu, "Rays: A ray searching method for hard-label adversarial attack," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1739–1747, 2020.

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.



(a) Clean Image



(b) Adversarial Image

Figure 1: Image 0 of CIFAR10 Dataset

(a) Clean Image

(b) Adversarial Image

Figure 2: Image 1 of CIFAR10 Dataset

[6] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[7] MetaMain, "Vitrobust - evaluating vision transformer robustness." GitHub repository, 2024. `https://github.com/MetaMain/ViTRobust/tree/6e3f3d548b3c463128c14749c39229a6fed6da0a`.

[8] Anonymous, "Vision transformers and robustness to adversarial attacks." YouTube video, 2024. `https://www.youtube.com/watch?v=pcYoymda49c`.

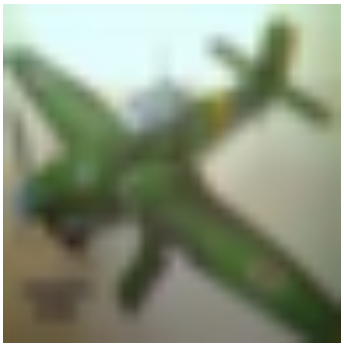
[9] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7838–7847, 2021.

[10] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," 2021.

(a) Clean Image

(b) Adversarial Image

Figure 3: Image 2 of CIFAR10 Dataset



(a) Clean Image

(b) Adversarial Image

Figure 4: Image 3 of CIFAR10 Dataset

(a) Clean Image

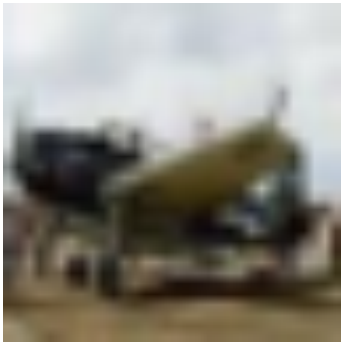
(b) Adversarial Image

Figure 5: Image 4 of CIFAR10 Dataset

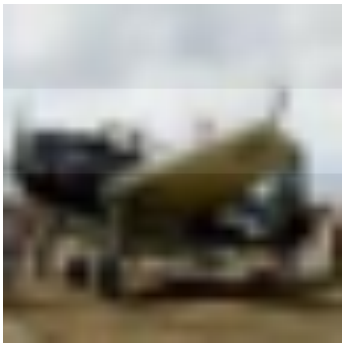
(a) Clean Image

(b) Adversarial Image

Figure 6: Image 5 of CIFAR10 Dataset



(a) Clean Image



(b) Adversarial Image

Figure 7: Image 6 of CIFAR10 Dataset

(a) Clean Image

(b) Adversarial Image

Figure 8: Image 7 of CIFAR10 Dataset

(a) Clean Image (b) Adversarial Image

Figure 9: Image 8 of CIFAR10 Dataset

(a) Clean Image (b) Adversarial Image

Figure 10: Image 9 of CIFAR10 Dataset