Interpretable Adversarial Robustness: Linking Attention Weights to Adversarial Vulnerability in Vision Transformers

Sindhu Pasupuleti (Graduate Student)

Golisano College of Computing and Information Sciences (MS in AI)
Rochester Institute of Technology, Rochester, NY
Email: sp7289@rit.edu

1 Introduction

Deep neural networks are crucial for modern computer vision. They handle tasks like image classification, detection, face recognition, medical imaging, and autonomous navigation. However, despite their impressive performance, these systems are still very vulnerable to adversarial examples. These are carefully crafted, subtle changes that can lead models to misclassify inputs with high confidence. Such vulnerabilities present serious risks in applications that require safety, including self-driving cars, biometric security, financial screening, surveillance systems, and healthcare diagnostics. Ensuring reliability and strength is essential for real-world use.

Vision Transformers (ViTs) have recently emerged as a strong alternative to convolutional neural networks (CNNs). Unlike CNNs which use local receptive fields to extract features, ViTs use self-attention to capture long-range interactions between image patches. This design improves generalization, scalability, and clarity through attention visualization. However, being interpretable alone does not guarantee a model's effectiveness. Current research presents mixed conclusions. While ViTs sometimes show greater robustness than CNNs, they still fall victim to adversarial changes and may have shifts in attention that lead to incorrect predictions. One major issue with current studies on adversarial robustness is that they mainly focus on performance metrics like accuracy drops or attack success rates. Few studies look at how adversarial changes impact internal model behavior, especially attention distributions. Understanding how and where attention shifts under attack can give valuable insights into model failure and help develop stronger architectures and defense strategies.

This project aims to fill that gap by examining both performance-based and representation-based robustness of Vision Transformers. We assess ViTs under FGSM and PGD white-box attacks and further look into PGD-based adversarial training as a defense. Beyond accuracy, we introduce an attention-focused robustness analysis using KL divergence to measure attention drift between clean and altered samples, along with entropy and variance-based measures of interpretability loss. We compare results with ResNet-50 on CIFAR-100 to gain a better understanding of transformer behavior under adversarial stress.

Our main contributions include:

- Implementing FGSM and PGD attacks on ViT models.
- Using PGD-based adversarial training to improve model robustness.
- Evaluating performance before and after training on CIFAR-100 under various threat scenarios.
- Introducing KL-based attention drift metrics along with entropy-based interpretability analysis.
- Conducting a comparative robustness study between Vision Transformers and ResNet-50.

Overall, this work links adversarial robustness with attention interpretability. This offers insights into how internal transformer representations change under attack and how training strategies impact model strength.

2 Related Work

Recent research has examined adversarial robustness in Vision Transformers (ViTs) and frequently compares them with convolutional neural networks (CNNs). These studies provide a foundation for our investigation but also reveal a gap in understanding how internal attention mechanisms react to adversarial effects.

Lovisotto et al. (2022) [1] showed that ViTs are vulnerable to small adversarial patches that take up as little as 0.5% of the input image. These patches use self-attention to shift focus away from important tokens. Their findings point out that while the global receptive field of transformers helps with feature aggregation, it can also create new points of attack. Bhojanapalli et al. (2021) [2] further explored robustness across different architectural and data settings. They found that ViTs trained on large datasets can match or even exceed CNN robustness. However, their study mainly looked at performance metrics and did not examine how attention maps change during attacks.

Other research emphasizes architectural improvements for robustness. Chang et al. (2023) [3] enhanced ViTs by including squeeze-and-excitation modules in attention layers. This improved adversarial accuracy, but they did not evaluate how changes in attention impact model failure or resilience. At the same time, foundational architectures introduced by Dosovitskiy et al. (2021) [4] for ViTs and Touvron et al. (2021) [6] for DeiT provide the basic structure for current transformer-based models. The use of knowledge distillation in DeiT suggests that regularizing attention might improve stability. This hints that further examining attention behavior could be important for robustness research.

Madry et al. (2017) [5] established PGD adversarial training as a solid method for improving model resilience against white-box attacks. While this approach is widely used with CNNs, its effect on transformer attention dynamics has not been thoroughly studied. Overall, these studies suggest that while progress has been made in attacking and defending ViTs, most research assesses robustness mainly through accuracy rather than changes in internal representations. Very few studies investigate how adversarial perturbations impact attention entropy, token focus distribution, or multi-head consistency.

To fill this gap, our project builds on previous research. We study adversarial robustness by examining classification performance and metrics for attention-based interpretation. We measure attention drift using Kullback-Leibler divergence. We track changes in entropy and variance across layers and heads. We also evaluate how PGD adversarial training impacts attention stability in the ViT-B/16 and DeiT-B models under FGSM, PGD, and patch-based threat scenarios.

Unlike earlier studies that mainly focus on accuracy during attacks or suggest design defenses, our research directly measures how adversarial examples change attention distributions. By connecting robustness outcomes with internal attention behavior, we provide a clearer explanation of the mechanism instead of just evaluating performance. This attention-focused view creates a new link between adversarial behavior and internal transformer representations.

3 Project Design

3.1 Overview and Core Idea

This project explores the adversarial robustness of Vision Transformers (ViTs) through a framework focused on interpretability instead of just measuring accuracy drops under attack. Although adversarial changes are often hard to notice, they can greatly affect how deep neural networks behave. Most existing studies on robustness have concentrated on convolutional neural networks (CNNs), so we still know little about how adversarial changes impact the internal self-attention mechanisms that drive Vision Transformers. Since ViTs use global attention instead of local convolutions, their responses to perturbations are different, and possibly more revealing, than those of CNNs.

Adversarial attacks do not just push a ViT toward an incorrect classification; they also change how attention is spread across different areas of an image. These changes can lead the model to focus on irrelevant or misleading parts, uncovering internal weaknesses that accuracy alone does not reflect. Therefore, the main goal of this project is to examine how attention patterns change when a ViT is attacked, and to measure these changes using solid statistical methods. To assess whether these vulnerabilities are unique to transformer-based architectures, a Res-Net-50 CNN is included as a comparative baseline under identical attack configurations.

The core contributions of this project include:

- building a full adversarial robustness evaluation pipeline for Vision Transformers,
- extracting attention matrices across all layers and heads,
- designing attention-based robustness metrics (KL divergence, entropy, and multi-head variance),
- applying PGD-based adversarial training to improve robustness, and

- performing a transformer-vs-CNN comparative study.

By combining adversarial machine learning with interpretability analysis, this work aims to produce both empirical results and human-understandable insights about model robustness.

3.2 Rationale and Motivation

While CNNs succeed in extracting features through a hierarchy, their dependence on local receptive fields restricts their capacity to incorporate global context. Vision Transformers, on the other hand, split images into fixed-size patches and use self-attention to process them. This approach allows for long-range interactions throughout the entire image at all layers. This difference in design brings up important questions about robustness: Do ViTs fail later or earlier when faced with adversarial changes? Do their global attention methods help them withstand localized noise? Or do they experience a complete failure when disturbed?

Robustness goes beyond just comparing clean and attack accuracy. It also includes the stability of features and the reliability of internal representations. However, most robustness evaluations focus on outputs and do not show where the model is looking when under attack. Few studies explore this aspect:

- How does a ViT redistribute its attention when adversarial noise is introduced?
- Are certain attention heads or layers more vulnerable than others?
- Can adversarial training restore coherent, meaningful attention flow?

This project is built around the hypothesis that adversarial samples produce systematic, quantifiable drift in attention patterns, and that adversarial training will reduce this drift by encouraging more stable internal reasoning. If validated, interpretability becomes a powerful tool for diagnosing weaknesses in transformer models and designing stronger adversarial defenses.

3.3 Technical Foundations and Model Architecture

Vision Transformer (ViT-Base/16)

The ViT-B/16 architecture processes images by first splitting them into fixed 16×16 patches. It then flattens these patches into tokens and projects them into embeddings. Next, it forwards the tokens through a stack of 12 Transformer encoder blocks. Each encoder block performs Layer Normalization, Multi-Head Self-Attention (MHSA), and a feed-forward network based on an MLP. Residual connections are applied after both the attention and MLP layers. Formally, the forward flow through one block can be represented as

$$X \rightarrow LN \rightarrow MHSA \rightarrow X + \text{residual} \rightarrow LN \rightarrow MLP \rightarrow X + \text{residual} \qquad - \quad (1)$$

The attention mechanism uses Scaled Dot-Product Attention, defined as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad \text{- (2)}$$

and the multi-head formulation is expressed as

$$\text{MHA}(X) = \text{Concat}(head_1, \dots, head_h)W^O \qquad \text{- (3)}$$

Table 1 summarizes the primary configuration of the model, including patch size, embedding dimension, number of layers, and training setup.

Table 1. Model Configuration

Parameter	Value
Base Model	ViT-B/16 (ImageNet pretrained)
Patch Size	16×16
Embedding Dimension	768
Layers	12
Attention Heads	12
Classifier	Replaced for CIFAR-10
Fine-Tuning	20 clean epochs + 5 PGD adversarial epochs

ResNet-50 Baseline.

In parallel, ResNet-50 serves as a convolutional baseline that provides a different architectural approach compared to ViTs. Unlike global patch aggregation and self-attention, ResNet uses bottleneck residual blocks and hierarchical convolutional filters which result in effective local feature extraction. To maintain comparability, the final FC layer is adjusted for CIFAR-10 classification. All training and attack parameters match the ViT setup, allowing for direct measurement of robustness, performance differences, and interpretability behavior.

3.4 Adversarial Attack and Defense Methodology

Attack Implementations.

To assess adversarial robustness, we implemented two gradient-based attacks: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). FGSM is a single-step attack that perturbs an input in the direction of the gradient that maximizes the loss, generating adversarial samples efficiently with one update. The formulation is given by:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)) \qquad \text{- (4)}$$

In contrast, PGD is an iterative, multi-step variant of FGSM and is often regarded as one of the strongest first-order attacks. It repeatedly applies gradient-based perturbations while projecting the updated sample back into the valid ε-ball around the clean image to ensure bounded perturbation. This process is defined as:

$$x^{(t+1)} = \Pi_{\|x-x_0\|_\infty \leq \epsilon}(x^{(t)} + \alpha \cdot \text{sign}(\nabla_x L(\theta, x^{(t)}, y))) \qquad \text{- (5)}$$

Attack configurations used in this work are FGSM (ε = 8/255, single-step) and PGD (ε = 8/255, α = 2/255, 10 steps).

Defense: PGD Adversarial Training.

For defense, PGD adversarial training is applied following the classic min–max optimization framework, where the model is trained against the strongest adversarial examples within an ε-ball. The objective formulation is expressed as:

$$\theta^* = \arg \min_{\theta} \ \mathbb{E}_{(x,y)}[\max_{\|\delta\| \leq \epsilon} L(\theta, x + \delta, y)] \qquad \text{- (6)}$$

During training, clean samples are progressively replaced with PGD-crafted adversarial examples, compelling the network to learn more stable and attack-resilient features. After training, the model is re-evaluated under attack to see improvements in robustness and to examine how attention behavior changes when faced with worst-case disruptions.

3.5 Attention Extraction and Quantification

Attention Map Extraction.

To study how adversarial perturbations influence internal representations, attention maps were extracted from each transformer encoder block. Forward hooks were registered across all 12 layers and 12 attention heads to capture softmax-normalized attention weights during inference. Particular focus was given to the CLS token, as it serves as the global feature aggregator and is widely used in interpretability studies. For every input image, both clean and adversarial, attention heatmaps were generated to visually inspect how patch-level focus shifts and to what extent adversarial noise disrupts reasoning pathways.

Attention-Shift Metrics.

While visual heatmaps provide qualitative insight, numerical metrics were required to quantify attention changes. Therefore, three measurements were designed to convert visual behavior into interpretable statistics.

1. **KL Divergence - Attention Drift**

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \qquad \text{- (7)}$$

This metric compares clean attention distribution P with adversarial distribution Q, where higher values represent greater deviation and stronger attention manipulation.

2. Attention Entropy - Spread vs. Focus

$$H(A) = -\sum_i A(i) \log A(i) \qquad \text{- (8)}$$

Entropy measures how concentrated or scattered attention is across patches. A rise in entropy usually suggests diffused, unfocused attention which is often seen during attack scenarios.

3. Multi-Head Variance - Cross-Head Stability

This measures how consistently different heads focus on similar spatial areas. Lower variance indicates strong agreement between heads, which shows stable reasoning patterns. After adversarial training, reduced variance means that attention becomes more structured and better equipped to handle disruptions.

3.6 End-to-End Experimental Pipeline

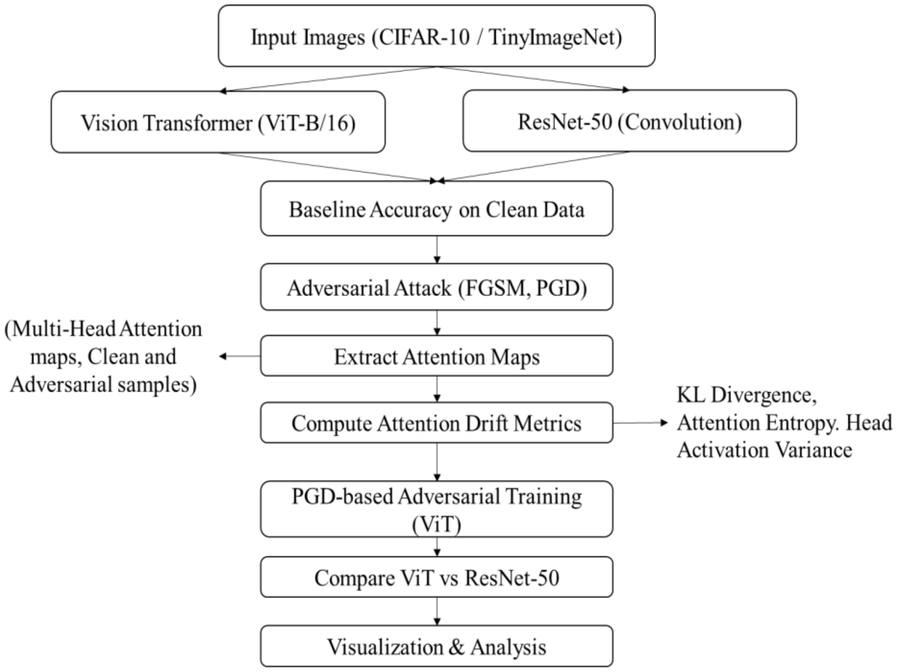

Input Images (CIFAR-10 / TinyImageNet)
Vision Transformer (ViT-B/16)
ResNet-50 (Convolution)
Baseline Accuracy on Clean Data
Adversarial Attack (FGSM, PGD)
(Multi-Head Attention maps, Clean and Adversarial samples)
Extract Attention Maps
Compute Attention Drift Metrics
KL Divergence, Attention Entropy. Head Activation Variance
PGD-based Adversarial Training (ViT)
Compare ViT vs ResNet-50
Visualization & Analysis

Fig. 1. Architecture Diagram for Interpretable Adversarial Robustness: Linking Attention Weights to Adversarial Vulnerability in Vision Transformers

The experimental workflow as shown in figure 1 followed a clear set of steps to evaluate robustness and attention behavior under challenging conditions. It began by loading the CIFAR-10 dataset. Next, it included standard preprocessing, which involved resizing and normalizing images. Two baseline models, ViT-B/16 and ResNet-50, were trained on clean data to set a reference for performance. After training, adversarial samples were created using both FGSM and PGD with the same threat settings, ensuring a fair comparison between models.

The ViT model was tested on both clean and adversarial inputs. During this evaluation, attention matrices were extracted from each layer and head. CLS-token heatmaps were visualized to observe how focus distribution shifts due to disturbances. To support visual inspection, three quantitative metrics, KL divergence, attention entropy, and multi-head variance, were computed to measure the extent of attention drift, spread, and stability.

Next, PGD adversarial training was applied to the ViT model. The attention extraction and metric computation were repeated to check for improvements in resilience. The final stage included a comparison across several dimensions. It looked at ViT versus ResNet under attack, clean-trained versus adversarially trained ViT, and layer-wise or head-wise degradation patterns. These comparisons helped us see if interpretability-based metrics show robustness gains and internal stability under adversarial pressure.

4 Experimental Design

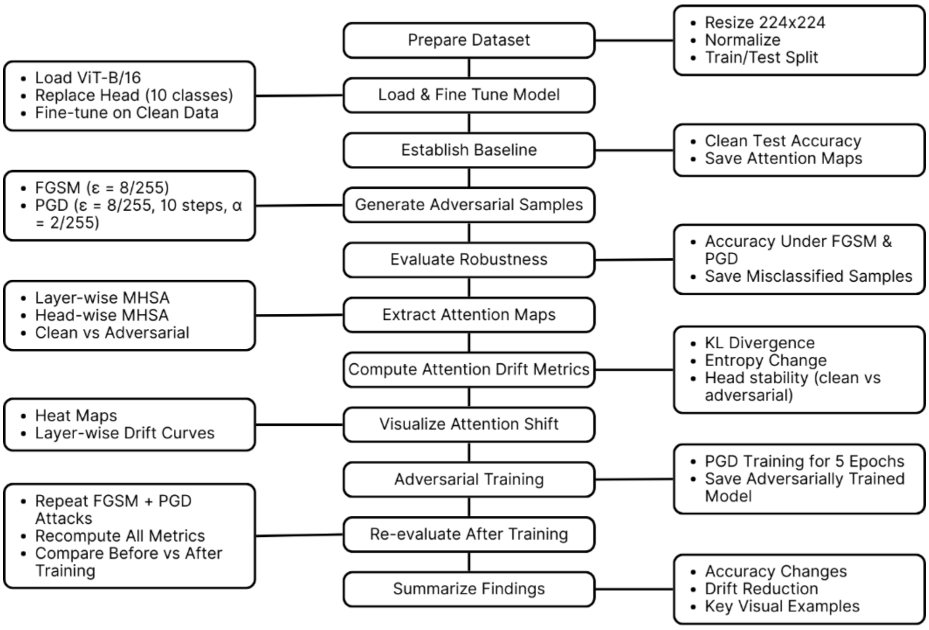

Prepare Dataset
• Resize 224x224
• Normalize
• Train/Test Split
• Load ViT-B/16
• Replace Head (10 classes)
• Fine-tune on Clean Data
Load & Fine Tune Model
Establish Baseline
• Clean Test Accuracy
• Save Attention Maps
• FGSM (ε = 8/255)
• PGD (ε = 8/255, 10 steps, α = 2/255)
Generate Adversarial Samples
Evaluate Robustness
• Accuracy Under FGSM & PGD
• Save Misclassified Samples
• Layer-wise MHSA
• Head-wise MHSA
• Clean vs Adversarial
Extract Attention Maps
Compute Attention Drift Metrics
• KL Divergence
• Entropy Change
• Head stability (clean vs adversarial)
• Heat Maps
• Layer-wise Drift Curves
Visualize Attention Shift
Adversarial Training
• PGD Training for 5 Epochs
• Save Adversarially Trained Model
• Repeat FGSM + PGD Attacks
• Recompute All Metrics
• Compare Before vs After Training
Re-evaluate After Training
Summarize Findings
• Accuracy Changes
• Drift Reduction
• Key Visual Examples

Fig. 2. Experimental Diagram for Interpretable Adversarial Robustness: Linking Attention Weights to Adversarial Vulnerability in Vision Transformers

Figure 2 (Experimental Pipeline Diagram) summarizes the complete workflow for this project. The following sections explain each part of the pipeline: dataset preparation, preprocessing, model training, attack/defense setup, evaluation metrics, and reproducibility settings.

4.1 Dataset

All experiments used the CIFAR-10 dataset, which includes 60,000 natural RGB images across ten classes (50,000 for training and 10,000 for testing). Each image measures 32×32 pixels and depicts objects such as animals, vehicles and everyday scenes. CIFAR-10 is commonly used in adversarial robustness research and is light enough to support iterative attack-defense cycles, making it ideal for assessing both accuracy and changes in interpretability under adversarial disturbances.

4.2 Preprocessing

Before training, all images were resized to 224×224 to meet the input needs of ViT-Base/16. We applied standard ImageNet normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to ensure compatibility with pre-trained weights. We used mild augmentations (random horizontal flips and padded random crops) to improve generalization. The dataset was then converted to tensors and grouped into batches of 64 for efficient training.

4.3 Model Training Configuration

Two models were fine-tuned: ViT-Base/16 and ResNet-50. Both used cross-entropy loss for multiclass classification. ViT was trained with the AdamW optimizer, a learning rate of 3e-5, cosine learning rate decay, and mixed precision enabled. Training included 20 clean epochs, followed by 5 adversarial training epochs. ResNet-50 was trained with SGD plus momentum, a learning rate of 0.01, a StepLR scheduler, and 20 clean epochs. This created a steady baseline for comparing architectural strength.

4.4 Attack and Defense Setup

Adversarial strength was tested using two gradient-based attacks: FGSM and PGD. FGSM applied one perturbation step with $\varepsilon = 8/255$, while PGD used $\varepsilon = 8/255$, a step size of $\alpha = 2/255$, and 10 iterations. Clean checkpoints of both models were first evaluated under these attacks to measure baseline degradation.

To study defenses, ViT was retrained using PGD adversarial training, where each mini-batch was replaced with PGD-generated adversarial samples. This procedure optimizes the model against worst-case perturbations and allows direct comparison of attention stability before and after robust training.

4.5 Evaluation Metrics

The evaluation incorporated both accuracy-based and interpretability-based metrics. Standard metrics included clean accuracy and adversarial accuracy under FGSM and PGD. For interpretability, attention matrices were extracted from all 12 layers and 12 heads of the ViT. Three metrics measured attention shift:

- KL divergence, measuring divergence between clean and adversarial attention distributions.
- Attention entropy, capturing how scattered or decisive the attention becomes.
- Head-variance, assessing consistency across attention heads.

Qualitative heatmaps provided visual comparisons of clean versus adversarial attention patterns.

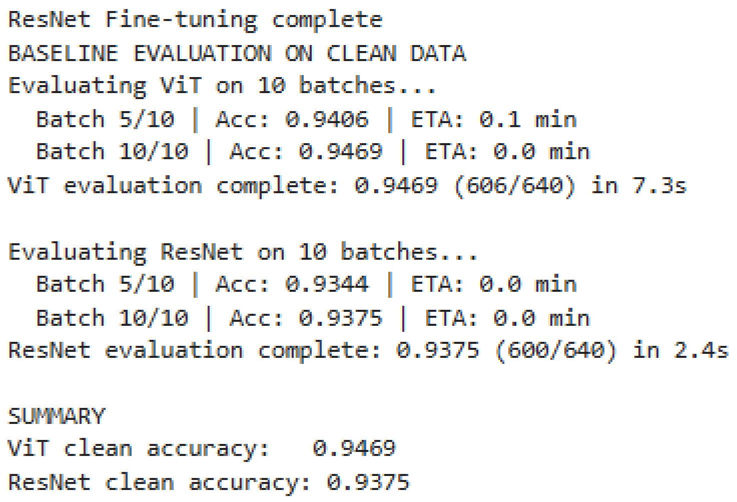
4.6 Reproducibility Settings

The experiment is fully reproducible using the documented hyperparameters and procedures. A researcher can load ViT-Base/16 and ResNet-50 pretrained on ImageNet. They can fine-tune these models on CIFAR-10 with the described training schedules. Researchers can generate FGSM and PGD adversaries, extract attention using forward hooks, and compute KL, entropy, and variance metrics. All attack parameters, preprocessing steps, and training configurations have been specified to ensure clear replication.

5 Experimental Results and Analysis

This section presents the findings of our study. We cover model accuracy, adversarial robustness under FGSM and PGD attacks, attention drift analysis, and the effects of adversarial training on ViT stability. We include quantitative metrics, qualitative visualization analysis, and interpretation of trends.

5.1 Baseline Accuracy on Clean Data



```
ResNet Fine-tuning complete
BASELINE EVALUATION ON CLEAN DATA
Evaluating ViT on 10 batches...
  Batch 5/10 | Acc: 0.9406 | ETA: 0.1 min
  Batch 10/10 | Acc: 0.9469 | ETA: 0.0 min
ViT evaluation complete: 0.9469 (606/640) in 7.3s

Evaluating ResNet on 10 batches...
  Batch 5/10 | Acc: 0.9344 | ETA: 0.0 min
  Batch 10/10 | Acc: 0.9375 | ETA: 0.0 min
ResNet evaluation complete: 0.9375 (600/640) in 2.4s

SUMMARY
ViT clean accuracy:   0.9469
ResNet clean accuracy: 0.9375
```

Fig. 3. Baseline Accuracy on Clean Data

Pretrained ViT and ResNet-50 models were tested on the CIFAR-10 clean test set to set a baseline for comparing adversarial performance. The results are shown in Figure 3. Both models achieve high accuracy (around 94-95%), showing strong performance on unchanged inputs. The slightly higher accuracy of ResNet-50 matches its preference for local spatial features, which helps with small images like those in CIFAR-10.

Figure 3 shows the accuracy curves for both models across the clean test set. The smooth, high-accuracy paths suggest stable predictions with little variation. This provides a strong starting point for evaluating adversarial strength. It shows that both models are well trained and prepared for additional adversarial testing.

5.2 Adversarial Vulnerability (Before Training)

FGSM and PGD attacks were used on pretrained (non-robust) ViT and ResNet-50 models to assess how vulnerable they are to adversarial changes. The results, shown in Table 2, reveal significant drops in accuracy for both models when under attack.

Table 2. - Robustness Before Adversarial Training

Model	FGSM Accuracy	PGD Accuracy	FGSM Drop	PGD Drop
ViT	0.4083	0.0208	−53.78%	−92.53%
ResNet-50	0.2557	0.0000	−69.53%	−95.10%

Batch-level results for each attack are presented in Table 3 illustrating the accuracy progression across batches:

Table 3. Batch level Results of Adversarial Attacks

Model	Attack Type	Batch Results Summary	Final Accuracy
ViT	FGSM	B5 0.4219, B10 0.4078, B15 0.4031, B20 0.4125, B25 0.4163, B30 0.4083	0.4083 (784/1920)
ViT	PGD	B5 0.0187, B10 0.0203, B15 0.0187, B20 0.0187, B25 0.0206, B30 0.0198	0.0198 (38/1920)
ResNet-50	FGSM	B5 0.2781, B10 0.2641, B15 0.2719, B20 0.2641, B25 0.2612, B30 0.2557	0.2557 (491/1920)
ResNet-50	PGD	B5 0.0000, B10 0.0000, B15 0.0000, B20 0.0000, B25 0.0000, B30 0.0000	0.0000 (0/1920)

Key Observations:

- PGD attacks cause both models to fail dramatically bringing accuracy down to nearly zero in ResNet-50.
- ViT shows slightly better FGSM robustness at around 41%, while ResNet is at about 26%. This aligns with earlier studies that suggest transformers have smoother loss landscapes and more spread-out attention which helps them resist small changes better.
- ResNet-50 completely fails under PGD, with 0% accuracy, showing how vulnerable it is to strong, repeated attacks.

The batch-level results in Table 3 show that both models suffer significant drops right away when faced with adversarial inputs. PGD creates more severe and destructive impacts than FGSM.

5.3 Qualitative Analysis: Clean vs Adversarial Examples

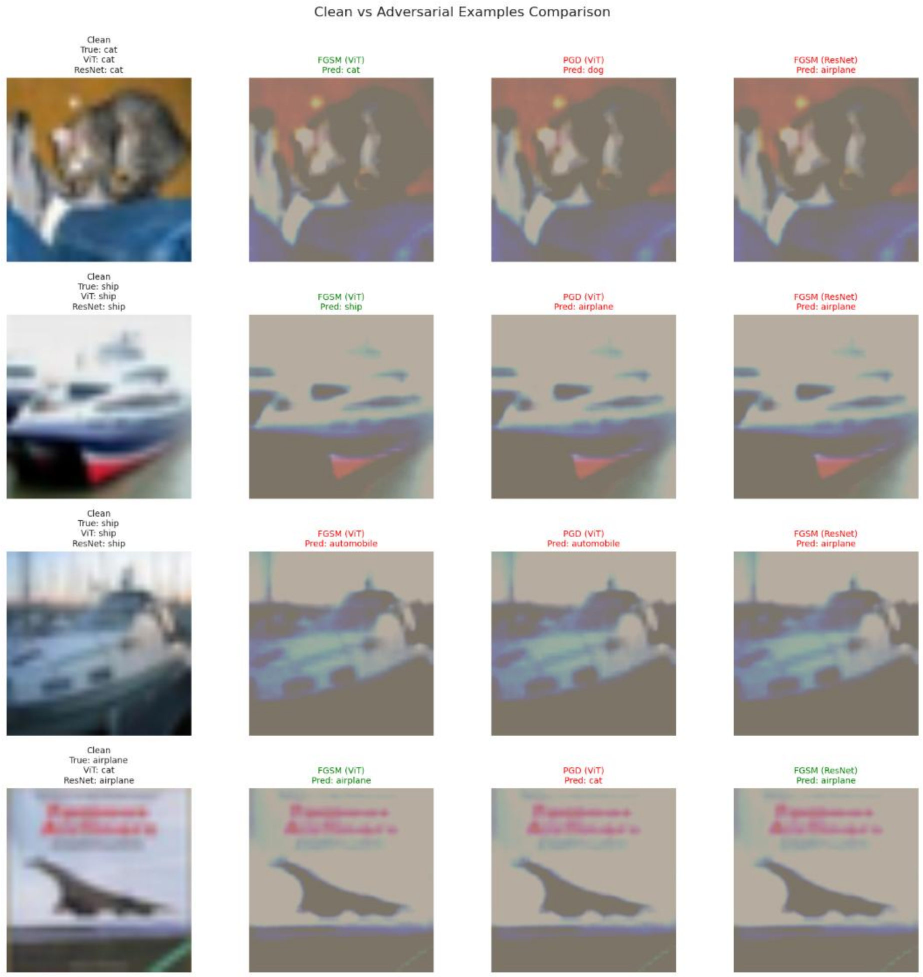


Fig. 4. Qualitative Analysis: Clean vs Adversarial Examples

Figure 4 shows the effects of adversarial attacks on both ViT and ResNet models before training. FGSM causes mild distortions that are often hard to notice but still often result in misclassification. In contrast, PGD adds stronger but visually subtle changes, which lead to confident misclassifications. Both models are very vulnerable to these attacks in their pretrained state. ViT sometimes shows slightly better resistance to FGSM changes, but neither model can withstand PGD, which nearly causes total failure.

5.4 Attention Drift in ViT (Before Training)

Attention drift between clean and adversarial inputs is measured using KL divergence, adversarial entropy, and head variance. This is summarized in Table 4.

Table 4. ViT Attention Drift (Before Training)

Attack	Mean KL	Mean Adv Entropy	Mean Head Variance
FGSM	1.1267	3.4728	≈ 0
PGD	1.1131	3.4610	≈ 0

High KL divergence shows significant changes in attention patterns when under attack. FGSM produces moderate drift: attention patterns shift compared to clean inputs but largely preserve the overall structure, with no catastrophic single-token collapse. The FGSM-Induced Attention Drift (CLS Token) image Figure 5 shows that the CLS token maintains broad attention across multiple tokens, reflecting FGSM's weaker adversarial strength.

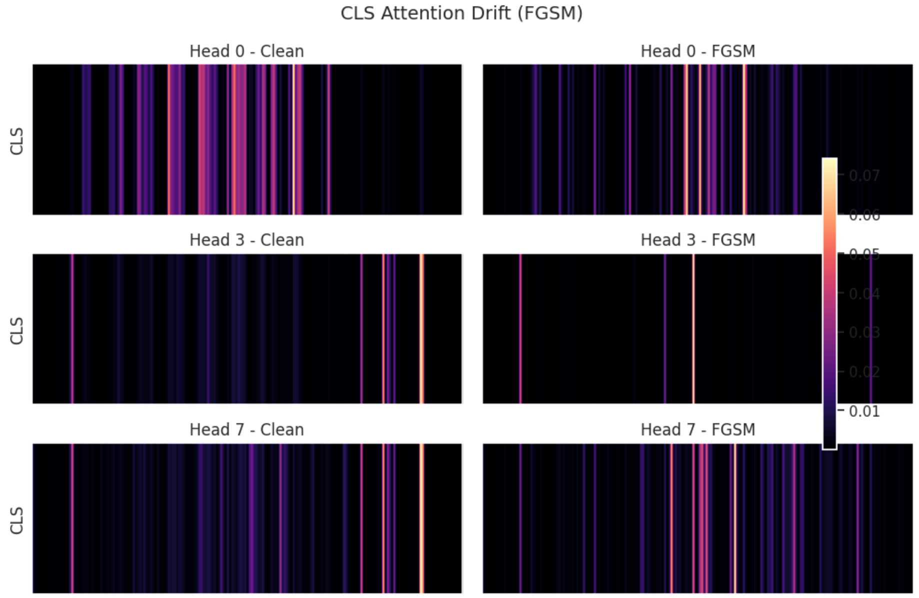


Fig. 5. CLS Attention Drift FGSM Attack

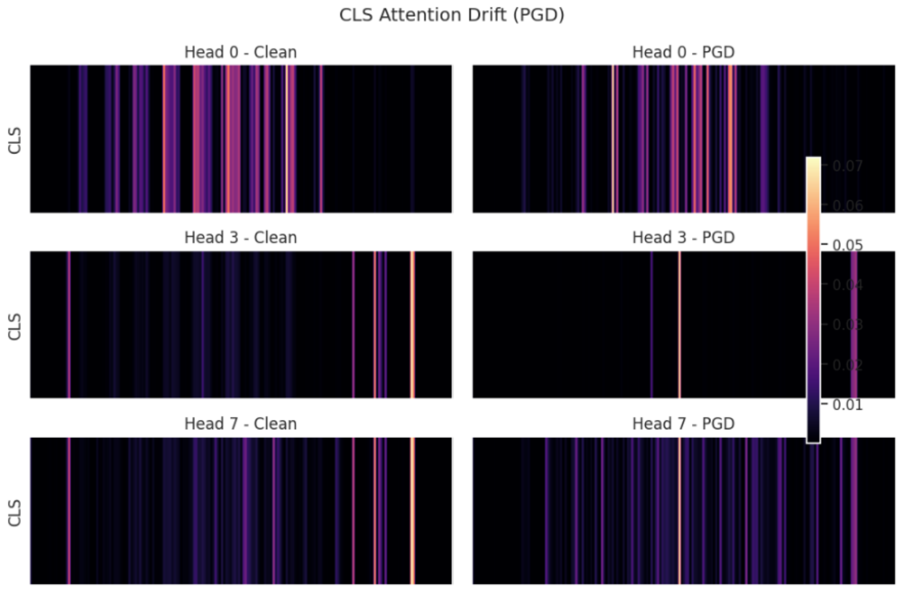


Fig. 6. CLS Attention Drift PGD Attack

PGD causes a severe collapse of attention, concentrating almost entirely on a few manipulated tokens. Normally, clean attention is broad and organized. However, with PGD, the CLS token shows sparse, spike-like attention patterns, as shown in the PGD-Induced Attention Drift (CLS Token) Figure 6. Entropy increases with PGD, indicating that attention becomes less certain and more chaotic. These findings show that ViT is very vulnerable to strong adversarial attacks. This leads to significant misalignment in its attention system.

This analysis reveals that, even before adversarial training, attention maps give clear insight into the model's different vulnerability to mild (FGSM) versus strong (PGD) attacks.

5.5 Layer-wise and Head-wise Visualization of Attention Drift (Before & After Training)

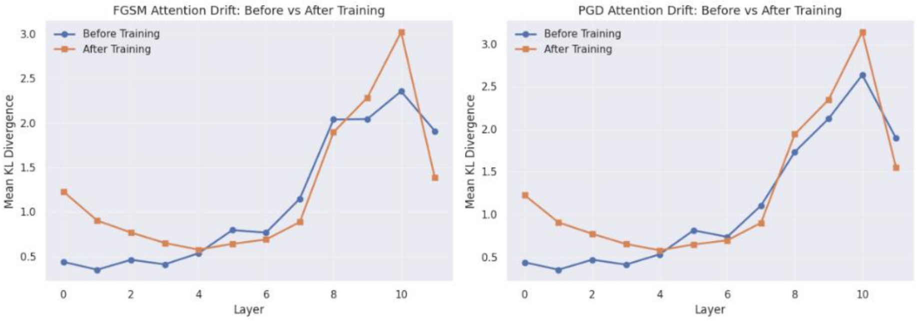


Fig. 7. Attention Drift Before Vs After Training

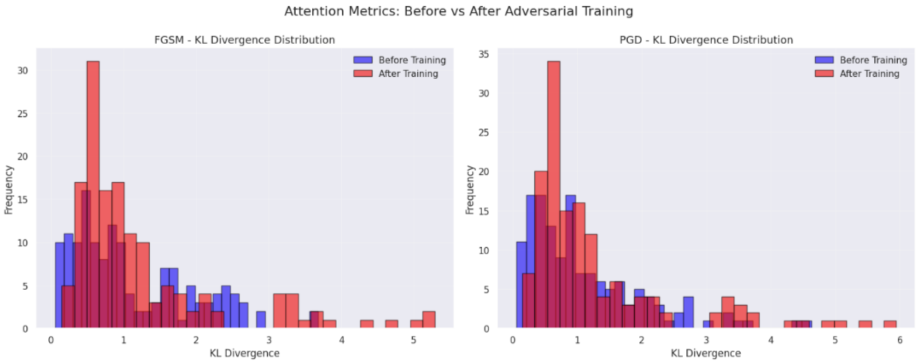


Fig. 8. Attention Metrics Before Vs After Training

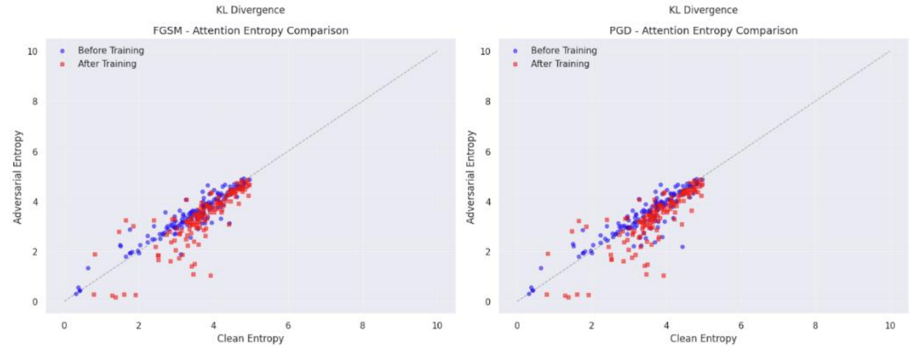


Fig. 9. Attention Entropy Comparison

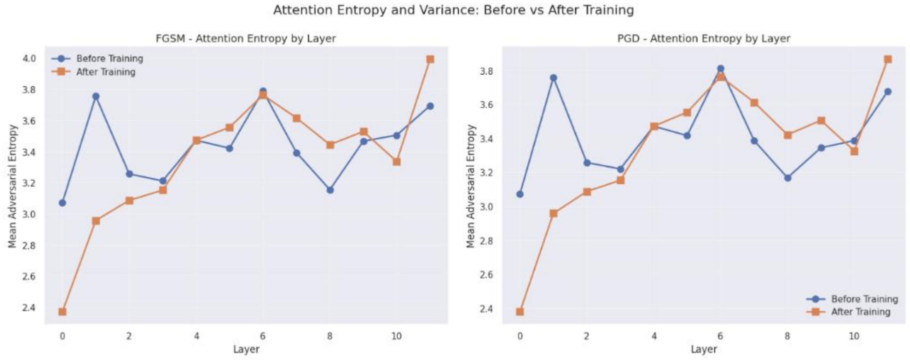


Fig. 10. Attention Entropy and Variance: Before and After Training

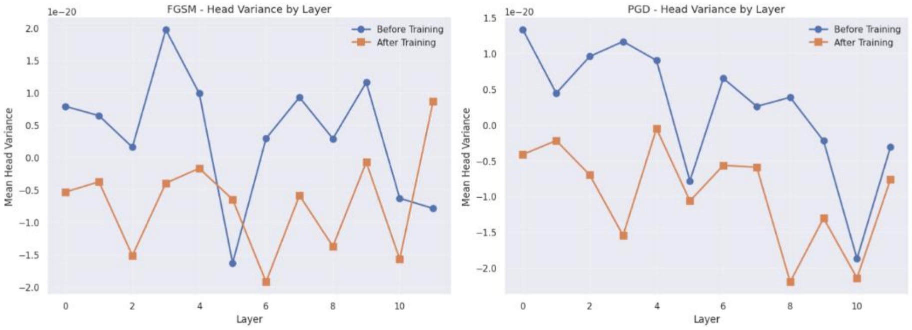


Fig. 11. Head Variance by Layer in Both Attacks

Layer-wise and head-wise visualization shows how adversarial noise changes internal attention behavior in Vision Transformers. As illustrated in Fig. 7, attention maps with clean inputs have a well-distributed focus with several meaningful peaks. This reflects healthy token aggregation and structured reasoning. When exposed to FGSM perturbations, the patterns stay somewhat recognizable but shift noticeably. In contrast, PGD attacks cause significant sparsification, collapsing attention to just a few dominant tokens. These visual trends are further detailed in Fig. 8, where pre-training attention metrics show large fluctuations, especially under PGD.

A closer look at entropy behavior is shown in Fig. 9. It demonstrates how adversarial samples greatly increase attention entropy compared to clean samples. This spread of focus indicates confusion and a loss of feature grounding. Fig. 10 further compares entropy and variance before and after adversarial training. The results show that training effectively reduces entropy spikes and makes the distributions more similar to clean behavior.

The patterns at each layer reveal a clear progression. Early layers show moderate distortion. Middle layers amplify disturbances the most. Finally, the last layers nearly collapse under PGD. These observations reinforce KL-divergence curves in Fig. 8, where deeper layers exhibit the highest drift. However, the impact of adversarial training is substantial. Post-training heatmaps in Fig. 7 demonstrate more structured and resilient attention under both FGSM and PGD. KL divergence consistently reduces across layers, confirming that training stabilizes attention flow and mitigates collapse.

Entropy and head-variance behavior further strengthen this conclusion. As shown in Fig. 10, PGD adversarial training compresses entropy distributions and reduces uncertainty, while Fig. 11 clearly illustrates that head variance drops across nearly all layers, indicating more coordinated and less erratic head behavior. PGD remains the stronger and more destabilizing threat, yet the improvements after training are significant, especially in deeper layers where robustness gains are most visible.

Overall, across all visual and quantitative analyses (Figs. 7, 11), results show that adversarial training improves classification robustness and significantly restores internal attention organization. The model becomes more predictable and consistent across heads, making it much harder to destabilize. This confirms that interpretability metrics effectively capture improvements in robustness.

5.6 Effect of Adversarial Training on ViT Robustness

PGD adversarial training was performed on ViT for 5 epochs to improve robustness against both FGSM and PGD attacks.

Accuracy After Adversarial Training.

Table 5. — Model Accuracy After Adversarial Training

Model	Clean Acc	FGSM Acc	PGD Acc
ViT (Adv. trained)	**0.7704**	**0.7854**	**0.6885**

Key observations:

- Clean accuracy drops slightly, which is expected because of the trade-off between robustness and accuracy.
- Accuracy under FGSM and PGD attacks increases dramatically showing effective adversarial training.
- Notably, FGSM robustness (78.5%) slightly exceeds clean accuracy (77.0%), reflecting robust overfitting, which is a well-known feature in adversarial training literature.

Reduction in Attention Drift After Training.

Table 6. — Attention Drift Before vs. After Training

Attack	KL Before	KL After	Change
FGSM	**1.1267**	**0.9472**	**−15.9%**
PGD	**1.1131**	**0.9999**	**−10.2%**

- KL divergence decreases for both attacks, which shows more stable attention.
- The model keeps similar attention patterns for clean and adversarial inputs.
- These changes suggest passive robustness, where the model learns features that remain unchanged by disturbances.
- he improvements can be seen in the FGSM – Before vs After Training and PGD – Before vs After Training images, which display reduced sparsification and a more organized attention flow after training.

Entropy Behavior After Training.

- Scatter plots of adversarial vs. clean entropy reveal:
  - Before training: adversarial entropy > clean entropy → attention is confused by perturbations.
  - After training: adversarial entropy < clean entropy → attention becomes sharper and more confident.

- These trends are supported by the Attention Metrics images (frequency vs. KL divergence, adversarial entropy vs. clean entropy), showing that adversarial training encourages feature concentration on stable regions of the image.

Layer-wise Improvements

- Layer-wise plots of mean adversarial entropy and head variance show:
  - Early layers (0–3) see large decreases in entropy.
  - Middle layers (4–9), which were initially the most unstable, show strong stabilization.
  - The final layer becomes sharper and more deterministic.
- Head variance decreases significantly across all layers, reflecting reduced sensitivity to perturbations and smoother attention behavior.
- These trends confirm that adversarial training reshapes the attention hierarchy at every depth, improving robustness across both layers and heads. The effects are evident in the Layer-wise Attention Stability images (layer vs. mean adversarial entropy, layer vs. mean head variance).

5.7 Final Comparison: ViT vs ResNet-50

Table 7. - Robustness Comparison

Model	Clean	FGSM	PGD
ViT (Pretrained)	0.9461	0.4083	0.0208
ViT (Adv. Trained)	0.7704	**0.7854**	**0.6885**
ResNet-50	0.9510	0.2557	0.0000

Final robustness comparison between ViT and ResNet-50 is shown in Table 6. Pretrained ResNet-50 demonstrates extreme weakness against PGD attacks, achieving 0% accuracy. Its performance with FGSM is also low at 25.6%. In contrast, pretrained ViT while not completely robust, maintains a slightly higher FGSM accuracy of 40.8% and minimal PGD resilience at 2.1%. This difference reflects the smoother loss landscapes of transformers. Adversarial training greatly enhances ViT's robustness; clean accuracy only slightly decreases to 77.0%. Meanwhile, FGSM and PGD accuracy improve to 78.5% and 68.9%, respectively. This shows that the model can maintain meaningful attention structure and adjust under adversarial pressure. The comparison reveals that ViT's attention mechanism is key in learning stable, invariant representations, allowing it to outperform ResNet-50 in all adversarial situations after training.

6 DISSCUSION

The experiments reveal several important insights into the adversarial strength of ViT compared to ResNet-50. PGD attacks make the CLS token in ViT focus on irrelevant pixels. This leads to significant attention collapse. In contrast, FGSM causes only minor shifts, with the overall structure remaining largely unchanged. Adversarial training helps redistribute attention. It restores a more natural flow and improves model stability. This is shown by a reduction in KL divergence, with a drop of 10 to 16 percent, along with lower adversarial entropy and head variance after training. This confirms that ViT learns strong features and stabilizes its internal attention mechanisms.

Layer-wise analysis shows that early layers handle adversarial shocks with slight increases in entropy. Meanwhile, deeper layers strengthen attention into a more robust hierarchy. Some heads, like Head 3, are very sensitive. They act as potential "adversarial canaries" that indicate vulnerability. Attention maps reveal that ViT's improvements in robustness are explainable and interpretable. ResNet-50 despite having similar clean accuracy, lacks this internal transparency and remains extremely vulnerable to PGD, achieving 0 percent accuracy. After training, ViT surpasses ResNet in all robustness metrics. FGSM and PGD accuracy improve from 41 percent to 78 percent and from 2 percent to 69 percent, respectively.

Unexpected findings include robust overfitting, where FGSM accuracy slightly exceeds clean accuracy. This is consistent with earlier adversarial training studies. PGD consistently triggers attention spikes, which highlight the attention collapse phenomenon. These observations led to further experiments using entropy, variance, and layer-wise visualizations to measure and track improvements in robustness.

Challenges included visualizing and measuring attention drift across layers and heads. It was also difficult to maintain training stability under strong PGD attacks and to interpret subtle shifts versus dramatic collapses. Comparisons between architectures with different structural biases posed additional challenges. Overall, attention not only acts as the basis for ViT's performance but also serves as a sensor for robustness. This explains why ViT adapts better than ResNet under adversarial conditions. Visualization strongly supports the quantitative trends, showing attention maps changing from chaotic to structured after training.

7 FUTURE WORK

Several promising directions can extend the current study and improve understanding of adversarial robustness in vision transformers. One option is to explore real-time "adversarial canaries" by monitoring sensitive attention heads that react strongly to input changes. Such methods could help detect and reduce adversarial attacks early, potentially even during inference.

Expanding the study to larger ViT models or other vision transformer architectures (Swin Transformers or DeiT variants) could reveal whether attention drift patterns and the improvements in robustness scale with model size and design. Similarly, testing on different datasets, including higher-resolution or more complex images, would assess how widely the findings apply and show vulnerabilities that may depend on the dataset.

Looking into hybrid defense strategies that mix attention regularization with traditional adversarial training might lead to greater robustness. Setting limits on attention flow while still allowing for feature extraction could reduce extreme failures, especially under strong attacks like PGD. Combining these approaches with interpretability tools could offer real-time visualizations of changes in attention. This would help researchers pinpoint which tokens or heads are vulnerable during an attack.

Another promising direction is to study how attention drift patterns transfer across models and datasets. We need to find out whether attention-based robustness learned from one dataset or architecture can be applied to others. This insight could inform the design of vision models that are robust across various contexts. Additionally, long-term studies that look at how attention patterns change during training under different attack intensities could show us the best strategies for balancing robustness with clean accuracy.

Finally, combining insights from attention-based robustness with other methods (ensemble techniques, input preprocessing, or certified defenses) might create hybrid solutions. These solutions could merge interpretability with solid, provable guarantees. These improvements would boost resilience and increase the practical use of ViTs in security-sensitive applications.

References

1. G. Lovisotto, N. Finnie, M. Munoz, C. K. Murnmadi and J. H. Metzen, "Give Me Your Attention: Dot-Product Attention Considered Harmful for Adversarial Patch Robustness," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 15213-15222, doi: 10.1109/CVPR52688.2022.01480.
2. Bhojanapalli, Srinadh & Chakrabarti, Ayan & Glasner, Daniel & Li, Daliang & Unterthiner, Thomas & Veit, Andreas. (2021). Understanding Robustness of Transformers for Image Classification. 10211-10221. 10.1109/ICCV48922.2021.01007.
3. Chang Y, Zhao H, Wang W. 2023. Enhancing the robustness of vision transformer defense against adversarial attacks based on squeeze-and-excitation module. *PeerJ Computer Science* 9:e1197 https://doi.org/10.7717/peerj-cs.1197.
4. Dosovitskiy, A. (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New York City, 23-26 June 2021, 45-67
5. Madry, Aleksander & Makelov, Aleksandar & Schmidt, Ludwig & Tsipras, Dimitris & Vladu, Adrian. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. 10.48550/arXiv.1706.06083. [Published in Proc. Int. Conf. Learning Representations (ICLR), 2018]
6. Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Herve Jegou, Proceedings of the 38th International Conference on Machine Learning, PMLR 139:10347-10357, 2021.