Chennai Mathematical Institute

Explainability in Vision Transformers

Manami Das MDS202423

Riya Shyam Huddar MDS202431

A report presented for the course of
Applied Data Analysis
Under the guidance of
Prof. Y Chandramouli
Year: 2025

Contents

1 Acknowledgment

This project draws inspiration from the research paper *"Explainability and Evaluation of Vision Transformers: An In-Depth Experimental Study"* by Sédrick Stassin, Valentin Corduant, Sidi Ahmed Mahmoudi, and Xavier Siebert, as well as from Jacob Gil's extensive work on Vision Transformer interpretability. We gratefully acknowledge these authors for their valuable contributions, which provided the conceptual foundation for our analysis of explainability techniques in transformer-based vision models.

We also express our sincere thanks to Prof. Y. Chandramouli for his guidance and for suggesting research directions in Explainable AI. His support and feedback were instrumental throughout the development and refinement of this project.

2 Work Contribution

Contributor	**Contribution Summary**
Manami Das	Developed the motivation section on CNN explainability and contributed to parts of the methodology for Vision Transformer explainability. Assisted with visualization design for both the PPT and the report.
Riya Shyam Huddar	Studied background literature and underlying algorithms. Implemented the Vision Transformer explainability methods. Collected supporting research papers and contributed content for the PPT and report.

Table 1: Contributions of Team Members

Both members thoroughly read and understood the full paper. The work in this report was completed collaboratively, involving shared problem-solving, discussions, and mutual support throughout the project.

Abstract

Vision Transformers (ViTs) have emerged as powerful alternatives to CNNs due to their global self-attention mechanisms, but their decision-making process remains difficult to interpret. This project explores and evaluates explainability techniques tailored to ViTs, including Attention Rollout and Gradient Attention Rollout, and compares them with CNN-based methods such as Grad-CAM. To provide more objective evaluation, we introduce a foreground–background attention framework using Mask R-CNN, enabling a quantitative assessment of how well ViTs focus on meaningful image regions. Our analysis shows that gradient-based rollout produces sharper, class-specific, and more semantically aligned explanations, making it more suitable for interpreting ViT predictions. The study highlights the need for transformer-specific interpretability tools in safety-critical and high-stakes applications.

3 Introduction

Motivation

Deep learning has achieved remarkable success in computer vision tasks such as classification, detection, and segmentation. While Convolutional Neural Networks (CNNs) have long dominated these domains due to their strong inductive biases, Vision Transformers (ViTs) have recently emerged as powerful alternatives. By operating on image patches and using global self-attention, ViTs capture long-range dependencies more effectively than CNNs.

However, the same flexibility that gives ViTs strong performance also makes them more difficult to interpret. Unlike CNNs, whose hierarchical spatial features naturally lend themselves to intuitive tools such as Grad-CAM, ViTs lack explicit locality and distribute information globally across patches. As a result, traditional CNN-based interpretability methods often fail or produce misleading insights when applied to transformer architectures.

In real-world, high-stakes applications—such as medical imaging or autonomous driving—an accurate but opaque model is insufficient. Understanding *why* a ViT makes a prediction is essential for trust, safety, and bias detection. This motivates the development of interpretability methods that reveal how ViTs allocate attention and which patches influence their decisions.

Problem

Despite their strong performance, ViTs remain challenging to interpret due to their global attention mechanisms. This raises several key questions:

- **How does a ViT process an image across layers?**
- **Which patches are most influential for the final prediction?**
- **How can we quantify attention on foreground versus background regions?**

- **How do attention patterns change with different target classes?**

The core problem addressed in this work is to design and evaluate techniques that answer these questions through reliable, intuitive, and class-sensitive explanations of Vision Transformer behavior.

Goal of the Project

The goal of this project is to analyze and compare explainability techniques tailored to Vision Transformers. We aim to identify influential patches, examine foreground–background attention distribution, and study class-specific attention shifts. Through this, we seek to make ViT decision-making more transparent, interpretable, and trustworthy.

4 Literature Survey

Vision Transformers (ViTs) were introduced by Dosovitskiy et al. [3], demonstrating that transformer architectures can outperform conventional convolutional networks on large-scale image recognition tasks. Rather than relying on convolutional feature hierarchies, ViTs treat an image as a sequence of fixed-size patches and use multi-head self-attention to model global dependencies. This design enables strong long-range reasoning but also complicates interpretability due to the lack of explicit spatial locality.

As ViTs gained adoption, the need to explain their decision-making processes became increasingly important. However, interpreting raw attention weights directly can be misleading and does not reliably represent the flow of information through the model. To address this, Abnar and Zuidema [2] proposed **Attention Rollout**, which aggregates attention across layers to approximate how information propagates from input patches to the final class token, forming the basis for many ViT explainability tools.

Although attention-based visualizations provide useful intuition, prior work has emphasized that attention alone is insufficient for faithful explanation. Chefer et al. [4] introduced a **gradient-based** attribution method for transformers that propagates relevance through both attention and MLP blocks, producing **class-specific** explanations that better align with model behavior.

Building on these ideas, we analyze attention fusion strategies and compare attention-based and gradient-based interpretability techniques using a pretrained DeiT-Tiny model. We further introduce a quantitative metric that enables systematic comparison of these explanation methods in terms of their semantic alignment and practical usability.

5 Methodology and Experimentation

5.1 Dataset

The experiments were conducted using a custom image dataset spanning varied object categories, together with selected ImageNet samples. All images were resized to 224×224

and normalized using standard ImageNet statistics.To keep evaluations in-distribution, images were aligned with ImageNet classes. The dataset includes foreground-dominant and background-dominant images, enabling analysis of how models allocate attention across different image regions. These images were used for evaluating multiple explainability methods:

- Grad-CAM for CNN explainability,
- Attention Rollout for Vision Transformers,
- Gradient Attention Rollout for Vision Transformers.

This combination of diverse image contexts supports a robust comparison between CNN-based and ViT-based interpretability methods.

5.2 CNN

To build an interpretability baseline for comparison with Vision Transformers, we used a pretrained ResNet-50 CNN and applied standard visualization techniques.

Methodology:

1. **Preprocessing:** Input images were resized, center-cropped, normalized using ImageNet statistics, and converted to tensors.
2. **Model:** A pretrained **ResNet-50** was used in evaluation mode for feature extraction and prediction.
3. **Inference:** For each image, logits were converted to probabilities and top-5 ImageNet classes were displayed.
4. **Feature Maps:** Intermediate convolutional layers were accessed through hooks to visualize how representations evolve from low-level edges to high-level object concepts.
5. **Explainability (Grad-CAM):** Grad-CAM was applied by capturing feature maps and gradients from a chosen layer and generating heatmaps that highlight influential regions in the image.

5.2.1 Experimentation

Experiments across multiple ResNet-50 layers show a clear progression of learned features:

1. **Layer 1:** captures edges, textures, and simple structures.
2. **Grad-CAM (Layer 1):** broad attention on the general object region.
3. **Layers 2–3:** represent parts, shapes, and mid-level semantics.
4. **Grad-CAM (Layer 3):** sharper focus on object-specific regions.
5. **Layer 4:** encodes high-level object features with abstract spatial patterns.
6. **Grad-CAM (Layer 4):** strongest localization on discriminative object regions.

Can Grad-CAM Explain Vision Transformers?

No. Grad-CAM relies on convolutional feature maps, whereas ViTs tokenize images into patches and mix them globally through self-attention, making gradients impossible to map to spatial locations. As shown by Chefer et al. (2021), Grad-CAM fails on transformers, so ViTs require dedicated methods such as **Attention Rollout** and **Gradient Attention Rollout**.

5.3 Vision Transformer

For explainability analysis, we use the DeiT-Tiny Vision Transformer, a lightweight ViT trained with knowledge distillation and designed to process images as sequences of patch tokens. The input image is resized to 224×224 and split into 16×16 patches, producing 196 patch embeddings that, together with a `[CLS]` token, form a sequence of 197 tokens. Each patch is flattened, projected through a linear embedding layer, and added with positional information before being processed through 12 Transformer blocks, each containing 3 multi-head self-attention (MSA) heads. At every layer, attention is modeled as a $3 \times 197 \times 197$ tensor, where tokens attend globally to one another. For interpretability, we later fuse the heads (using mean/max/min strategies) to obtain a single 197×197 attention map, making ViTs well suited for attention-based explanation compared to CNNs that rely on local receptive fields.

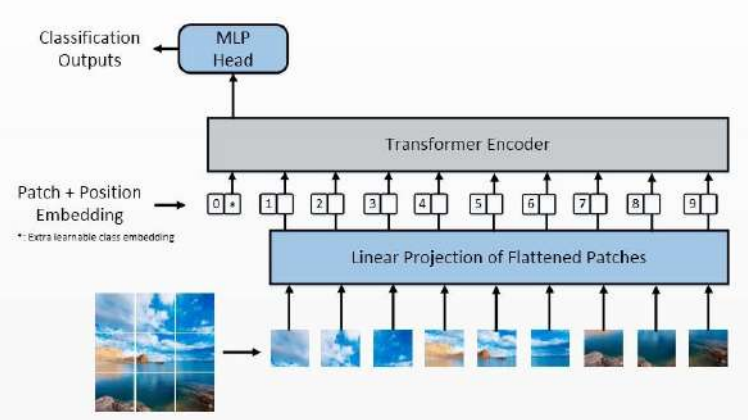


Figure 1: Vision Transformer (ViT) block diagram.

5.3.1 Explainability Methods

1. Attention Rollout Attention Rollout (Abnar & Zuidema, 2020) computes how information flows from input patches to the class token by recursively multiplying attention matrices across layers. Before this step, the three attention heads in each DeiT-Tiny block are fused (mean, max, or min) to obtain a single 197×197 matrix per layer. Let $A^{(l)}$ be the fused attention matrix at layer l. Residual connections are incorporated by adding an identity matrix:

$$\tilde{A}^{(l)} = \alpha A^{(l)} + (1-\alpha)I,$$

where $\alpha = 1$ by default. The final rollout matrix is computed as:

$$R = \tilde{A}^{(1)} \times \tilde{A}^{(2)} \times \cdots \times \tilde{A}^{(L)},$$

where $L = 12$ for DeiT-Tiny. The row in R corresponding to the class token indicates how much each patch contributed to the final prediction.

2. Gradient Attention Rollout Gradient Attention Rollout (Chefer et al., 2021) incorporates gradients to produce **class-specific** explanations. As in standard rollout, the three attention heads in each block are first fused (mean, max, or min) to obtain a single 197×197 attention matrix per layer. Given a fused matrix $A^{(l)}$ and its gradient $G^{(l)}$, class-relevant importance is computed as:

$$A^{(l)}_{\text{grad}} = \text{ReLU}\left(G^{(l)} \odot A^{(l)}\right),$$

where $\odot$ denotes element-wise multiplication. The class-specific rollout is then obtained by multiplying the L layerwise matrices:

$$R_{\text{grad}} = A^{(1)}_{\text{grad}} \times A^{(2)}_{\text{grad}} \times \cdots \times A^{(L)}_{\text{grad}},$$

with $L = 12$ for DeiT-Tiny. This produces an attention map that highlights only those patches contributing to the predicted class.

Max, Min, and Mean Attention Rollout

To analyze how ViTs distribute attention across patches, we compute the maximum, minimum, and mean of the final rollout vector R. Since each Transformer block produces three attention heads, these statistics provide a simple and consistent way to combine the information from all heads into a single interpretable measure of patch importance.

$$R_{\max} = \max_i R_i, \qquad R_{\min} = \min_i R_i, \qquad R_{\text{mean}} = \frac{1}{N}\sum_i R_i.$$

Interpretation:

- $R_{\max}$ highlights the most influential patch.
- $R_{\min}$ corresponds to the least relevant (typically background) region.
- R_{mean} reflects how broad or concentrated the overall attention is.

5.3.2 Experimentation

Experiments were conducted on the same dataset as the CNN baseline. For each image, we generated raw attention maps, Attention Rollout heatmaps illustrating global patch influence, and Gradient Attention Rollout heatmaps capturing class-specific relevance.

Observations:

- Attention Rollout provides a broad view of influential regions but often spreads attention across both foreground and background.
- Gradient Attention Rollout produces sharper, more localized maps by filtering out irrelevant attention paths.
- Attention Rollout tends to highlight the full semantic object, whereas Gradient Rollout focuses on the most discriminative parts (e.g., airplane nose or wings).
- These trends mirror our CNN findings: ViTs reason globally, while CNN Grad-CAM emphasizes spatially coherent regions.

Overall, the explainability maps show that ViTs distribute attention globally, while gradient-based rollout sharpens this distribution to reveal class-dependent reasoning. The contrast with CNN Grad-CAM highlights fundamental architectural differences in how both models extract and use visual features.

6 Novelty

Interpreting Vision Transformers has mostly relied on qualitative heatmaps obtained from attention rollout or gradient-based methods. While visually informative, such explanations do not quantify whether the model truly attends to the semantic object regions in the image. To address this gap, we introduce a simple, automated foreground–background evaluation framework that uses Mask R-CNN to estimate object masks, aligns them to the ViT patch grid, and measures the fraction of rollout attention that falls on foreground versus background. This provides a measurable, model-agnostic way to assess the faithfulness of ViT explanations.

6.1 Foreground Extraction via Mask R-CNN

We estimate the true object region using a pretrained Mask R-CNN model (COCO weights), which produces a binary foreground mask M^{FG}. The background mask is simply:

$$M^{BG} = 1 - M^{FG}.$$

This automated approach requires no manual annotation and works reliably across varied ImageNet images.

6.2 Patch-Level Alignment on the ViT Grid

The DeiT-Tiny ViT represents each image as a 14×14 grid of patches. We therefore downsample both masks to this grid. Let A_{ij} denote the attention (or gradient-weighted attention) assigned to patch (i, j). This allows direct comparison between model attention and semantic regions.

6.3 Quantitative Metric: Foreground–Background Attention Focus

To measure how much attention falls on the true object, we define:

$$FAF = \frac{\sum_{i,j} A_{ij}\, M_{ij}^{FG}}{\sum_{i,j} A_{ij}}, \qquad BAF = \frac{\sum_{i,j} A_{ij}\, M_{ij}^{BG}}{\sum_{i,j} A_{ij}}.$$

where - A_{ij}: attention score for patch (i, j), - M_{ij}^{FG}: 1 if the patch is part of the foreground, else 0, - M_{ij}^{BG}: 1 for background patches.

Since foreground and background partition the image, $FAF + BAF = 1$. A higher **FAF** indicates better focus on meaningful object regions; a higher **BAF** implies attention wasted on background.

Quantitative Results

We compute FAF/BAF for sample ImageNet images using both Gradient Attention Rollout and standard Attention Rollout. The side-by-side tables below summarize the results.

(a) Attention Rollout

Image	**FAF**	**BAF**
airliner	0.109	0.891
monastery	0.012	0.988
spaniel	0.481	0.519
telescope	0.294	0.706
gondola	0.543	0.457
shuttle	0.241	0.759
flute	0.922	0.078
bulbul	0.193	0.807
sweatshirt	0.448	0.552
piggy bank	0.467	0.533
gown	0.497	0.503

(b) Gradient Attention Rollout

Image	**FAF**	**BAF**
airliner	0.143	0.857
monastery	0.003	0.997
spaniel	0.703	0.297
telescope	0.448	0.552
gondola	0.497	0.503
shuttle	0.261	0.739
flute	0.927	0.073
bulbul	0.290	0.710
sweatshirt	0.672	0.328
piggy bank	0.573	0.427
gown	0.738	0.262

Table 2: Foreground (FAF) and Background (BAF) attention fractions for Attention Rollout (left) and Gradient Attention Rollout (right).

Gradient Attention Rollout achieves a higher mean FAF (0.478) compared to standard Attention Rollout (0.382), confirming that gradient-based explanations are more class-specific

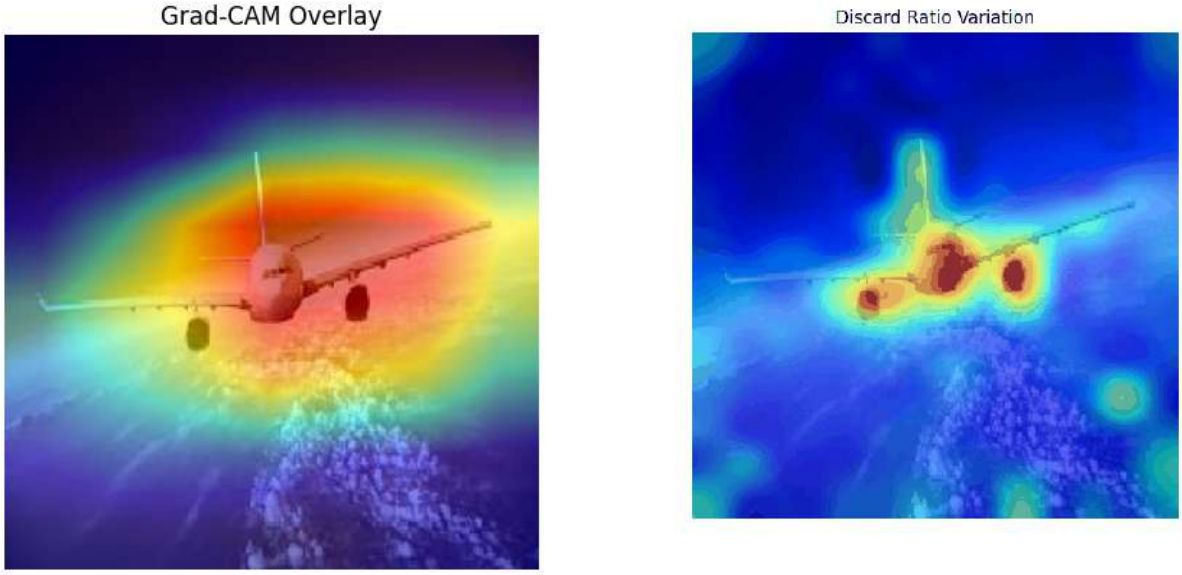
and better aligned with true object regions.

6.4 Why This is Useful

- **Quantitative evaluation:** Moves beyond visual inspection with measurable scores.
- **Detects model weaknesses:** High BAF reveals background-biased reasoning.
- **Benchmarking:** Enables fair comparison across explanation methods.
- **Fully automated:** No manual labels required; works across varied images.

This framework offers a concise yet powerful way to evaluate where Vision Transformers allocate attention and improves interpretation of ViT-based models.

7 Results



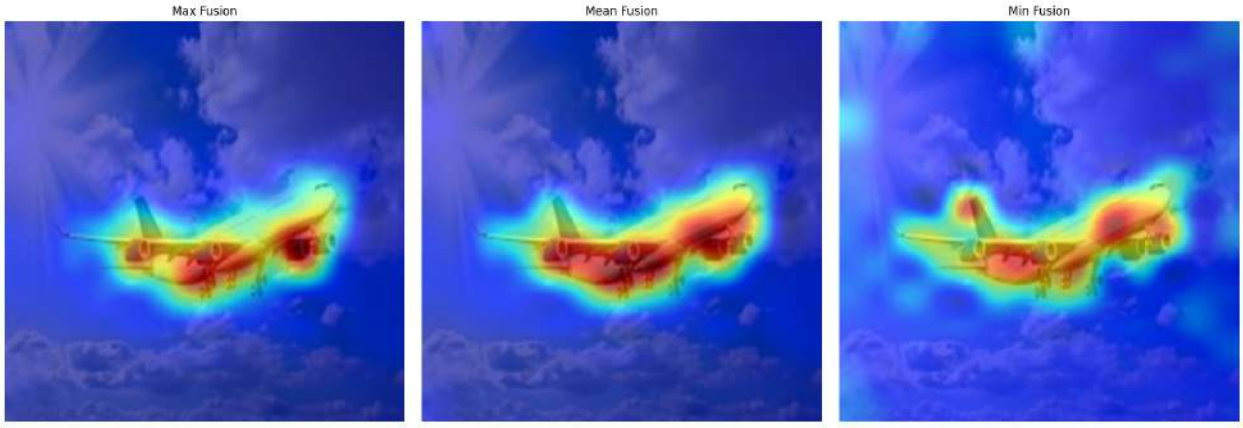
(a) CNN visualization using Grad-CAM

(b) ViT Attention Rollout Animation

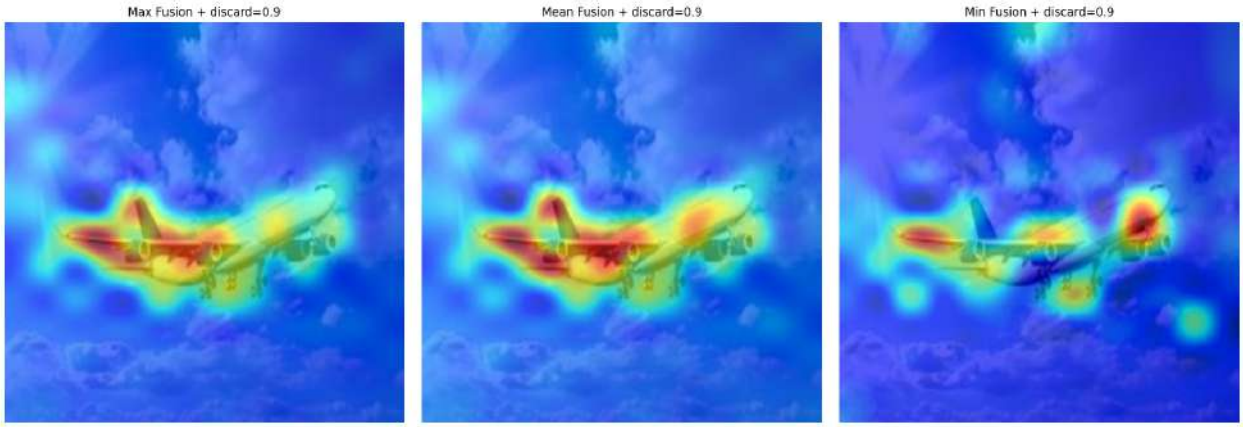
Figure 2: Comparison of CNN Grad-CAM and ViT Attention Rollout.

Interpretation: This figure contrasts CNN and ViT explanations. The Grad-CAM map (left) highlights compact, spatially localized regions, while the ViT Attention Rollout (right) distributes relevance more globally. This difference reflects the models' distinct feature-processing mechanisms and motivates our foreground–background analysis in later sections.

Min Mean Max fusion



(a) Max–Min–Mean Attention Rollout

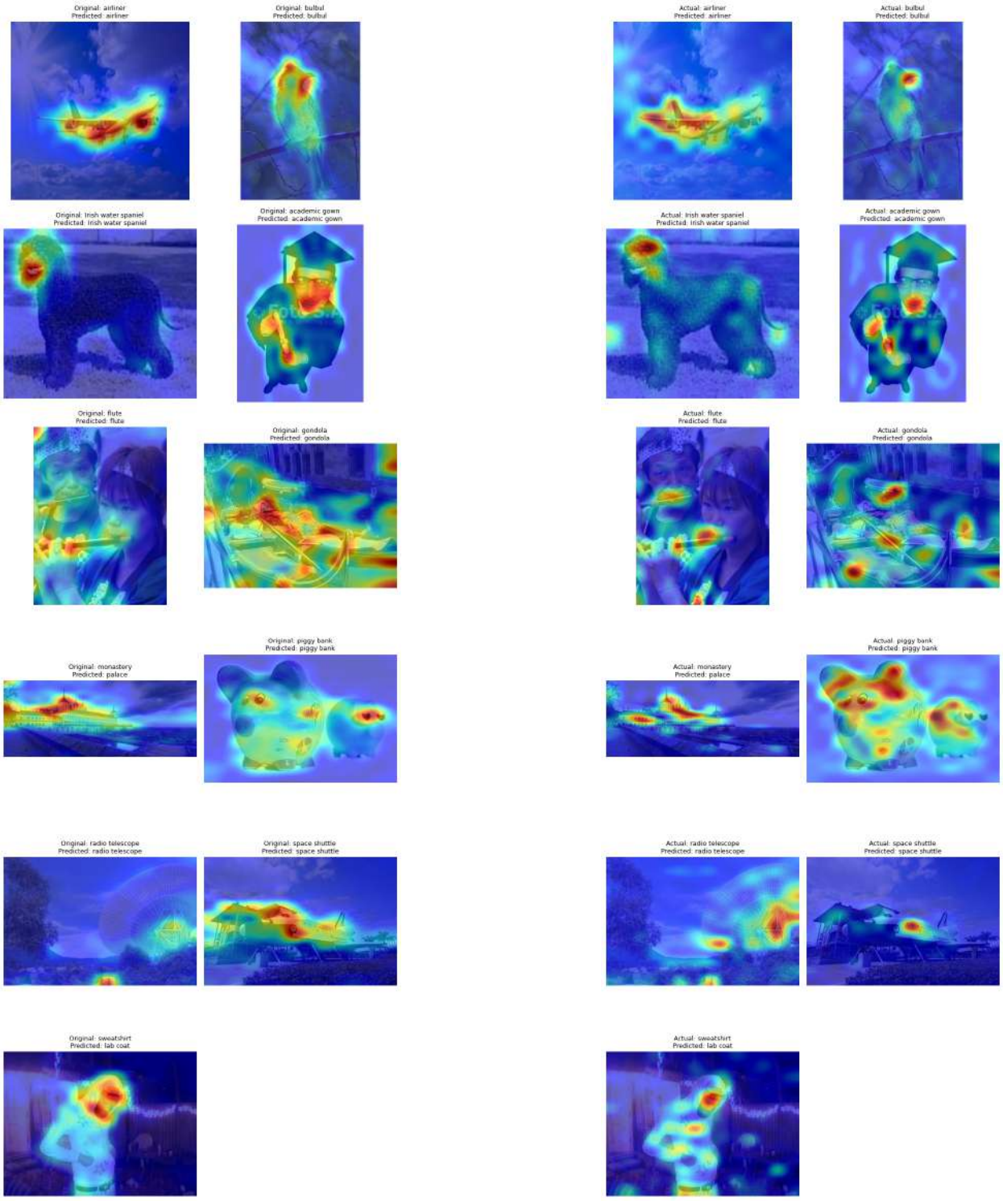


(b) Max–Min–Mean Gradient Attention Rollout

Figure 3: Comparison of rollout statistics for Attention Rollout and Gradient Attention Rollout.

Interpretation: The *max* fusion highlights the strongest response from any attention head, making it effective for identifying sharply discriminative patches. The *mean* fusion offers a more stable, consensus-based view of the model's focus across heads. The *min* fusion captures only regions consistently attended to by all heads, producing conservative but potentially under-sensitive explanations.

Attention Rollout vs Gradient Attention Rollout

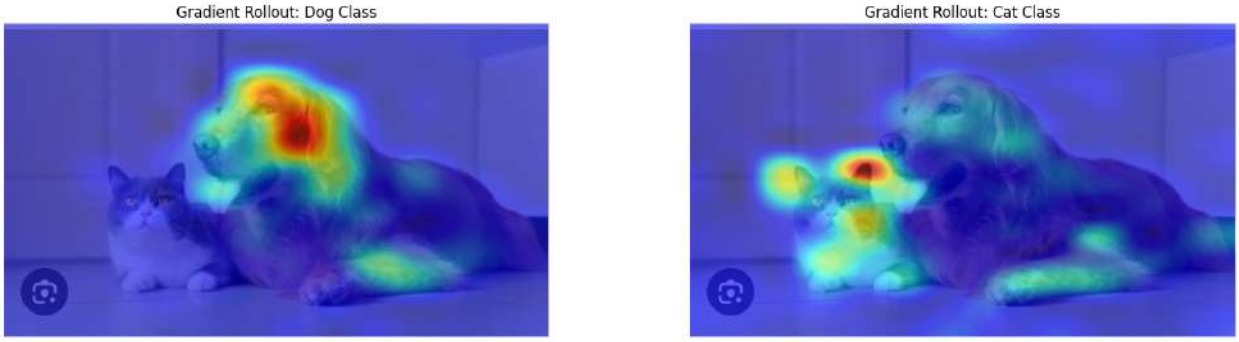


(a) Attention Rollout (b) Gradient Attention Rollout

Figure 4: Comparison of Attention Rollout and Gradient Attention Rollout across multiple images.

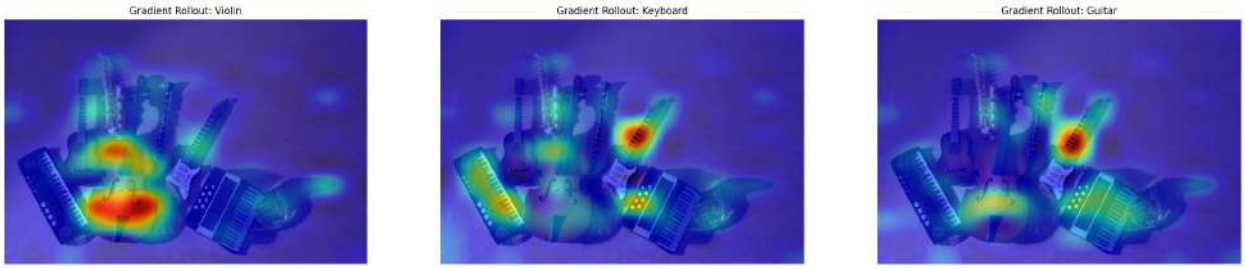
Interpretation: Across representative ImageNet samples, Attention Rollout tends to diffuse relevance over broad spatial regions, often including background areas. In contrast, Gradient Attention Rollout concentrates more sharply on class-relevant patches. These examples demonstrate how gradient signals refine raw attention into more precise and discriminative attribution maps.

Gradient Attention Rollout for different classes in same image



(a) Dog Class (b) Cat Class

Figure 5: Dog and Cat in the same image



(a) Violin (b) Keyboard (c) Guitar

Figure 6: Comparison of Violin, Keyboard, and Guitar classes.

Interpretation: Gradient Attention Rollout yields *class-specific* heatmaps, causing the highlighted regions to shift depending on the queried label. In the dog–cat example, the *dog* class focuses on the dog, while the *cat* class highlights only the cat. The same behaviour appears for violin, keyboard, and guitar, where each class activates the corresponding instrument. This shows that gradient-based rollout reliably isolates the regions most relevant to each predicted class.

8 Conclusion and Future Scope

- **CNN vs ViT explanations:** CNNs generate localized, spatially coherent saliency maps, whereas ViTs rely on global patch interactions, making traditional CNN-based explanation tools insufficient for transformers.
- **Attention Rollout vs Gradient Rollout:** Standard Attention Rollout produces broad but diffuse relevance, often highlighting background regions. Gradient Attention Rollout yields sharper, class-specific explanations due to the incorporation of gradient information.
- **FAF/BAF metric validation:** Our proposed Foreground Attention Fraction (FAF) and Background Attention Fraction (BAF) quantitatively confirm that Gradient Rollout aligns better with meaningful foreground regions, with a higher mean FAF (0.478) compared to standard Attention Rollout (0.382), demonstrating greater faithfulness and precision.
- **Contribution of this work:** The project moves beyond purely qualitative heatmaps by introducing a quantitative, automated foreground–background evaluation framework, enabling scalable and reproducible assessment of ViT interpretability.
- **Metric limitations and future improvement:** A more fine-grained evaluation using detailed semantic or part-based masks could improve precision, but generating such masks is computationally expensive and less practical at scale.
- **Scalability and generalization of the evaluation framework:** Although our quantitative experiments were conducted on a limited set of images, the pipeline itself is deliberately model-agnostic and easily scalable. Future work can extend the FAF/BAF metric to larger datasets, analyze class-specific trends, and incorporate alternative segmentation models to strengthen robustness.
- **Towards standardized benchmarking:** The proposed FAF/BAF approach provides a step toward establishing quantitative benchmarks for evaluating Vision Transformer explainability methods.
- **Interpretability disclaimer:** While our analysis uses attention as an interpretability tool, we do not equate attention with causal explanation; rather, our work probes ViTs to better understand what patterns they attend to and how these patterns relate to their outputs, recognizing that attention is only one component of the broader explanation puzzle.

Bibliography

[1] Jacob Gildenblat, *Exploring Explainability for Vision Transformers.* https://jacobgil.github.io/deeplearning/vision-transformer-explainability

[2] Abnar, S., and Zuidema, W., , Quantifying Attention Flow in Transformers., arXiv preprint arXiv:2005.00928, 2020. https://arxiv.org/abs/2005.00928

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. , *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.* arXiv preprint aarXiv:2010.11929, 2021. https://arxiv.org/abs/2010.11929

[4] Chefer, H., Gur, S., and Wolf, L., *Transformer Interpretability Beyond Attention Visualization.* https://pages.stat.wisc.edu/~bwu62/771/golub1996.pdf

[5] Sédrick Stassin, Valentin Corduant, Sidi Ahmed Mahmoudi, Xavier Siebert, *Explainability and Evaluation of Vision Transformers: An In-Depth Experimental Study*, https://www.researchgate.net/publication/377023845_Explainability_and_Evaluation_of_Vision_Transformers_An_In-Depth_Experimental_Study

[6] He, K., Gkioxari, G., Dollár, P., and Girshick, R., *Mask R-CNN*, arXiv preprint arXiv:1703.06870, 2017. https://arxiv.org/abs/1703.06870