Interpreting Vision Transformers

*K. Hebbar (6785768), C. Okafor (6781114), S. Banerjee (6790349)

Abstract

This study investigates the interpretability of Vision Transformers (ViT) [1], a compelling class of models employing transformer architectures to solve computer vision tasks. We explore diverse ViT variants, including base architectures with distinct patch sizes (16x16 and 32x32), varying transformer blocks and heads, and the DeiT [2], a modified version of ViT that leverages a teacher network.
Our approach involves calculating attention maps from attention matrices and visualizing these as self-attention maps to illustrate the model's attention progression across different transformer layers [3]. To enhance interpretability, we implement the attention rollout technique [4], offering a more comprehensive view of the model's focus across layers. Moreover, we introduce a gradient attention rollout, which incorporates gradients into attention computation, ensuring the attention maps reflect not only the model's focus but also regions contributing significantly to the final prediction [5].
Our visualizations reveal that the most insightful attention maps are generated by the DeiT and the vanilla ViT with a larger number of transformer layers and heads. These models capture more intricate patterns, probably due to their increased complexity and capacity. Our findings contribute to the ongoing efforts to enhance the interpretability of deep learning models [6] and foster trust in their predictions. Our contributions can be found here https://github.com/zuruoke/intrepreting-ViT.git.

1. Introduction

In the recent past, Vision Transformers (ViT) have emerged as a powerful approach for a variety of visual tasks, often outperforming traditional convolutional neural networks (CNNs) [1]. Inspired by the success of Transformer architecture in natural language processing (NLP), these models have shown promise in multiple image recognition tasks [2]. Yet, despite their impressive performance, the interpretability of ViTs remains a challenge, leaving the understanding of the factors influencing their predictions largely opaque. This impedes their adoption, especially in critical applications where understanding the model's decision process is essential [3].

In this study, our objective is to explore the

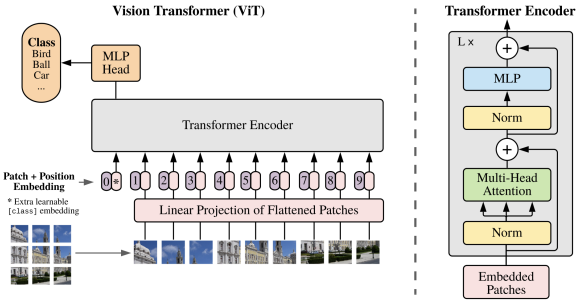


Fig 1: ViT model [1]

the intricacies of the ViTs by examining different variants of these models, including the base architecture with distinct patch sizes *(16x16 and 32x32)*, varying transformer blocks and heads, and the DeiT, a modified version of ViT [2].
Our analysis is anchored on pre-trained models, enabling a robust examination of these models' generalizable behaviour.
Our study focuses on computing and visualizing attention maps to understand the relevance of input features in Vision Transformers (ViTs). We generate score matrices, and using the attention and gradient attention rollout techniques, we track the model's attention evolution across layers. Notably, DeiT and larger ViTs provide the most insightful attention maps, suggesting their complex structure captures nuanced data patterns effectively. Specifically, DeiT's usage of a pre-trained ImageNet model enhances its focus on key image areas [2].

The main contributions of our study are summarized as follows:

1. We provide a detailed analysis of different ViT architectures, including base ViT and DeiT, using pre-trained models.
2. We compute and visualize self-attention maps for various ViT models, illustrating the evolution of attention across transformer layers.
3. We implement the attention rollout technique, offering a comprehensive, layer-wise view of the model's attention.
4. We introduce a gradient attention rollout that integrates gradients into the attention computation,

ensuring the produced attention maps reflect regions that significantly contribute to the final prediction.
5. We provide a detailed interpretation of the visualized attention maps, revealing insights into the workings of the ViT models and their focus areas.

Our study contributes to the ongoing efforts to improve the interpretability of deep learning models. By offering insights into the operation of ViTs, we hope to foster trust in these models and facilitate their wider adoption in domains where understanding the model's decisions is crucial.

2.0. Related Works

2.1. Vision Transformer Models

2.1.1. Vision Transformer (ViT)

Dosovitskiy et al. [1] introduced the Vision Transformer (ViT), a novel approach that applies the transformer architecture, originally designed for natural language processing, to computer vision tasks. In contrast to conventional Convolutional Neural Networks (CNNs) that handle images in their entirety, ViTs divide images into a sequence of patches, treating them as a 1D sequence. The self-attention mechanism of the transformers then captures the relationship among these patches, facilitating a global understanding of the image content.

2.1.2. Data-efficient Image Transformers (DeiT)

Touvron et al. [2] proposed a modified version of ViT, named Data-efficient Image Transformers (DeiT). DeiT is designed to be more data-efficient, requiring fewer data and computational resources compared to ViTs. DeiT leverages a technique called knowledge distillation, where it learns from a pre-trained teacher network, often an ImageNet model, to improve its performance. This process enables DeiT to make accurate predictions despite the data limitations.

2.2. Interpretability of Vision Transformers

2.2.1. VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers

Yu et al. [7] developed an interactive visualization tool, VL-InterpreT, for interpreting Vision-Language Transformers. This tool visualizes attention maps and the internal state of these models, providing insights into how they process and integrate multimodal inputs. This understanding is critical for debugging and improving these models.

2.2.2. Transformer Interpretability Beyond Attention Visualization

Vig et al. [8] delved deeper into the interpretability of transformer models beyond attention visualization. They argued that the insights gained from visualizing attention weights may not entirely capture the model's reasoning process. Thus, they proposed additional techniques, providing a more comprehensive understanding of the inner workings of these models.

2.3. Interpretability of Techniques for Visualising Attention Maps

2.3.1. Attention Rollout: Understanding and Visualizing Multi-head Attention

In the work of Jain and Wallace [9], they present an approach called "attention rollout" for interpreting and visualizing attention in multi-head transformer models. The attention rollout technique aims to provide a global perspective of the model's attention by aggregating the attention from all layers of the transformer.
In essence, attention rollout starts from the final layer of the transformer and sequentially rolls out the attention through the preceding layers. This technique takes into account the fact that the attention at each layer is conditioned on the attention of the previous layers, thus providing a more accurate and comprehensive visualization of the model's attention mechanism.
The authors demonstrated the effectiveness of this technique in providing interpretable insights into the model's decision-making process across a variety of NLP tasks. Their work suggests the potential of attention rollout for enhancing the interpretability of Vision Transformers as well.

2.3.2. Visualizing and Understanding Attention in Transformers

In the work of Vig et al. [10], they explore a suite of techniques to visualize, understand, and interpret attention in Transformer models. One key method they discuss is the aggregation of attention across layers and heads, which is closely related to the attention rollout technique.
They propose a method for decomposing attention to gain insights into the individual heads' behaviours and the specialized functions they might learn. This

involves constructing attention matrices and using dimensionality reduction techniques to visualize the structure of attention in these models.
The authors demonstrate their visualization techniques on a variety of tasks and datasets, showing how attention can reveal interesting structural properties of the input data and the model's decision-making process. This work provides valuable insights that can be applied to the interpretability of Vision Transformers and the attention rollout technique.

3. Methodology

Our study aims to interpret Vision Transformers (ViT) by analyzing the self-attention maps of various models. We conduct our investigation using pre-trained ViT models, including the vanilla ViT models and the ViT DeiT models.

3.1. Model Selection

The models selected for this study are:

1. Vanilla ViT models:
   - *vit_base_patch16_224*: Base ViT model with 16x16 patch size, 12 transformer layers, and 12 attention heads.
   - *vit_base_patch32_224*: Base ViT model with 32x32 patch size, 12 transformer layers, and 12 attention heads.
   - *vit_large_patch16_224*: Large ViT model with 16x16 patch size, 24 transformer layers, and 16 attention heads.
2. ViT DeiT model:
   - *deit_tiny_patch16_224*`: DeiT model with 16x16 patch size, 12 transformer layers, and 3 attention heads.

3.2. Image Preprocessing

Our preprocessing pipeline begins with resizing the input image to a *256x256* pixel resolution using a high-quality downsampling filter (order 3 interpolation). Subsequently, a central crop is performed, reducing the image dimensions to *224x224* pixels. The image is then transformed into a PyTorch tensor, a single data type multi-dimensional matrix, primed for subsequent computations.
In the final step, we normalize the tensor using mean and standard deviation values, specifically *0.485*, *0.456*, and *0.406* (mean) and *0.229*, *0.224*, and *0.225* (standard deviation) for the red, green, and blue channels, respectively. These values, derived from the ImageNet dataset, ensure the image tensor is optimally prepared for efficient processing by the Vision Transformer models.
Additionally, we conducted experiments with five distinct images to ensure the robustness and generalizability of our observations.

3.3. Calculating Attention Maps

To compute the attention maps, we start with the preprocessed image tensor. We extract corresponding attention matrices for the image via a PyTorch hook function, using a pre-trained Vision Transformer model. The hook function intercepts the *attn.attn_drop* module in each transformer layer, allowing us to capture the attention matrices.
The attention matrices, of shape *(h, n, n)*, where h is the number of attention heads and *n* is the attention matrix dimension, vary based on the image patch size.
For a *16x16* patch size, *n* is *197*, and for a *32x32* patch size, *n* is *50*. This dimension includes individual patches plus an additional classification *(cls)* token. The cls token, a learned representation, captures global information about the image.
Finally, we organize the attention matrices into a list of *length*, *l*, where *l* is the number of transformer layers in the model. This list provides a layer-wise attention distribution, giving insight into the model's attention patterns at various depths.

3.4. Computing Score Matrix

To investigate the attention distribution, we calculate a score matrix whose dimensions are based on the image's patch size. For *16x16* patches, we yield a *14x14* matrix, and for *32x32* patches, a *7x7* matrix.
The process begins by excluding the classification token from the attention matrix, producing an adjusted matrix, *A*. We then find the maximum value across each head in *A*, leading to a list of maximum values, *M*.
Next, we compute the mean of these maximum values, creating a matrix *C*. This step condenses the information from multiple heads into a single representation, showcasing the average peak attention value for each image patch.
Finally, we reshape *C* to produce the score matrix *S*, with each element indicating the model's attention intensity for a particular image patch. This process can be summarized by the equation:

$$S = reshape(mean(M), target\_dimensions) \quad (3.3)$$

Here, *reshape* denotes resizing the tensor, mean indicates calculating the mean values, and *M* represents the matrix of maximum values.

3.5. Implementing Attention Rollout

The Attention Rollout technique provides insights into the cumulative attention across layers in transformer-based models. It starts by fusing attention across all heads to form a fused attention matrix, *F*.
Subsequently, a portion of the lowest attention values are dropped from *F*, focusing on the most influential tokens, resulting in matrix *D*.
Next, *D* is combined with an identity matrix *I* and the sum divided by two, producing A, ensuring diagonal elements are higher than others.
Then, *A* is normalized across the sequence to obtain *N*, ensuring attention sums for each token equal one.
Finally, cumulative multiplication of *N* across all layers forms the rollout attention matrix *R*, showing total attention from the cls token to each specific token.
This process can be expressed as:

$$R = \Pi(\Sigma(D + I) / 2) \quad (3.4)$$

Here, Π stands for cumulative multiplication across all layers, Σ denotes normalization along the sequence, D represents the matrix after dropping low attention values, and *I* is the identity matrix.

3.6. Gradient Attention Rollout Technique

The Gradient Attention Rollout technique enhances understanding of attention distribution by integrating gradients of attention scores. Firstly, an identity matrix *I* is formed, and each attention matrix is fused with its gradient to form matrix *F*. Negative values in F are nullified, and a segment of low attention values is discarded to form *D*.
D is then combined with *I*, and the result is divided by two to form matrix *A*. *A* is normalized along the sequence to form *N*, ensuring attention sum for each token equals one.
Cumulative multiplication of *N* across layers gives the rollout attention matrix *R*. Elements in *R* relating to image patches are considered, reshaped to image size, and normalized to yield matrix *M* - the output of the Gradient Attention Rollout technique.
This process can be summarized as:

$$F = mean(A \circ G, axis = 1) \quad (3.4a)$$
$$F = max(F, 0) \quad (3.4b)$$
$$flat = top\_k(F, k = discard\_ratio) \quad (3.4c)$$
$$D = reshape(flat, shape(F)) \quad (3.4d)$$
$$M = \Pi(\Sigma((D + I) / 2)) \quad (3.4e)$$

Here, Π denotes cumulative multiplication, Σ normalization, *D* the matrix after discarding low attention values, and *I* the identity matrix. *M* represents the reshaped and normalized rollout attention matrix.

4. Experimental Results

In this section, we present the results derived from our proposed methodology, particularly emphasizing the outcomes of different visualization techniques employed.

4.1. Visualization of Progression in Self-Attention Maps

In this subsection, we display and discuss the transformation of self-attention maps across the various layers of the ViT model. The aim here is to demonstrate how the model progressively refines its attention, focusing on different regions of the input image, as the data traverses through the transformer layers. This is shown in *Fig 4.1*.

4.1.1. Observation and Findings

The heat maps from various Vision Transformer (ViT) models as showcased in *Fig 4.1* shows distinct attention distribution patterns.
The *vit_base_patch16_224* model concentrated attention in the final and a middle layer, with other layers showing a stable but diffused focus.
The *vit_base_patch32_224* model had the most attention in the last and initial layers, with a uniform distribution in the middle layers.
The *vit_large_patch16_224* model showed the highest heat in one of the final layers and the least in a middle layer, with earlier layers exhibiting a broad distribution.
Finally, the *deit_tiny_patch16_224* model concentrated attention in the final layer, with the rest of the layers presenting a more focused pattern.
These observations underline the unique dynamics of attention distribution across layers and models in the Vision Transformer architecture, revealing the complex ways these models process visual data.

4.2. Visualization of the Attention Rollout Technique

The Attention Rollout Technique is visualized in this part of our experimental results. This technique provides us with an aggregated view of the model's attention, offering a cumulative perspective over all transformer layers. The resulting visualizations help to understand how the model allocates and integrates

attention from a global standpoint. This is shown in *Fig 4.2*.

4.2.1. Observations and Findings

Observations drawn from implementing the Attention Rollout as shown in *Fig 4.2,* on the four Vision Transformer (ViT) models provide further insights into their interpretability.
Upon using a set of five images as input, the *deit_tiny_patch16_224* model consistently outperformed the others in generating attention maps that accurately captured the focus areas of the images.
The *vit_large_patch16_224* model followed closely, generating satisfactory results. Interestingly, the *vit_base_patch16_224* model seemed to struggle in identifying the focal areas of the images, indicating some differences in attention distribution compared to the other models.
This could be attributed to the unique feature of the *deit_tiny_patch16_224* model, which utilizes a teacher network during training to enhance its performance. This *knowledge distillation* approach might be responsible for the superior attention focusing ability of the *deit_tiny_patch16_224* model. It suggests that the presence of a guiding network during training can significantly influence the model's interpretability and attention distribution capabilities.

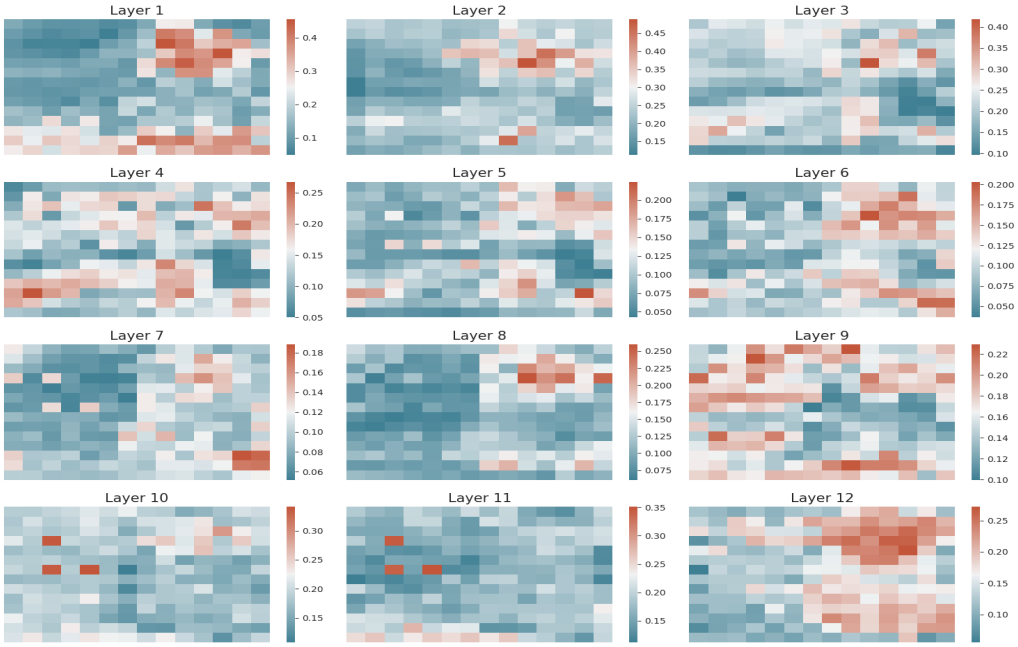
4.3. Visualization of the Gradient Attention Rollout Technique

Lastly, we present the visualizations resulting from the Gradient Attention Rollout Technique as shown in *Fig 4.3*. This technique enhances the attention rollout by incorporating gradients, which represent the influence each part of the input image has on the final prediction. The visuals derived from this technique are particularly insightful, as they highlight the image areas most critical to the model's decision-making process.

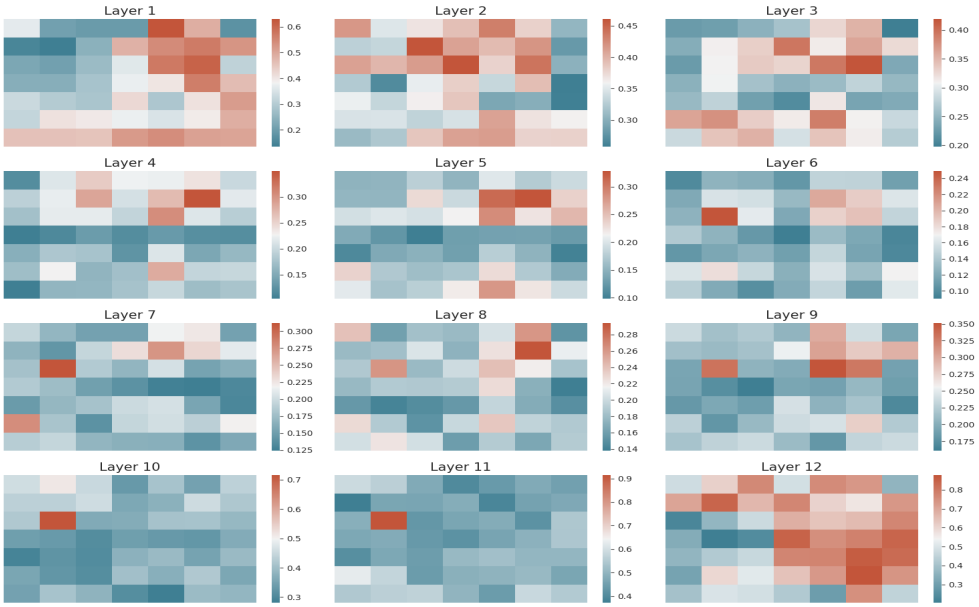
4.3.1. Observations and Findings

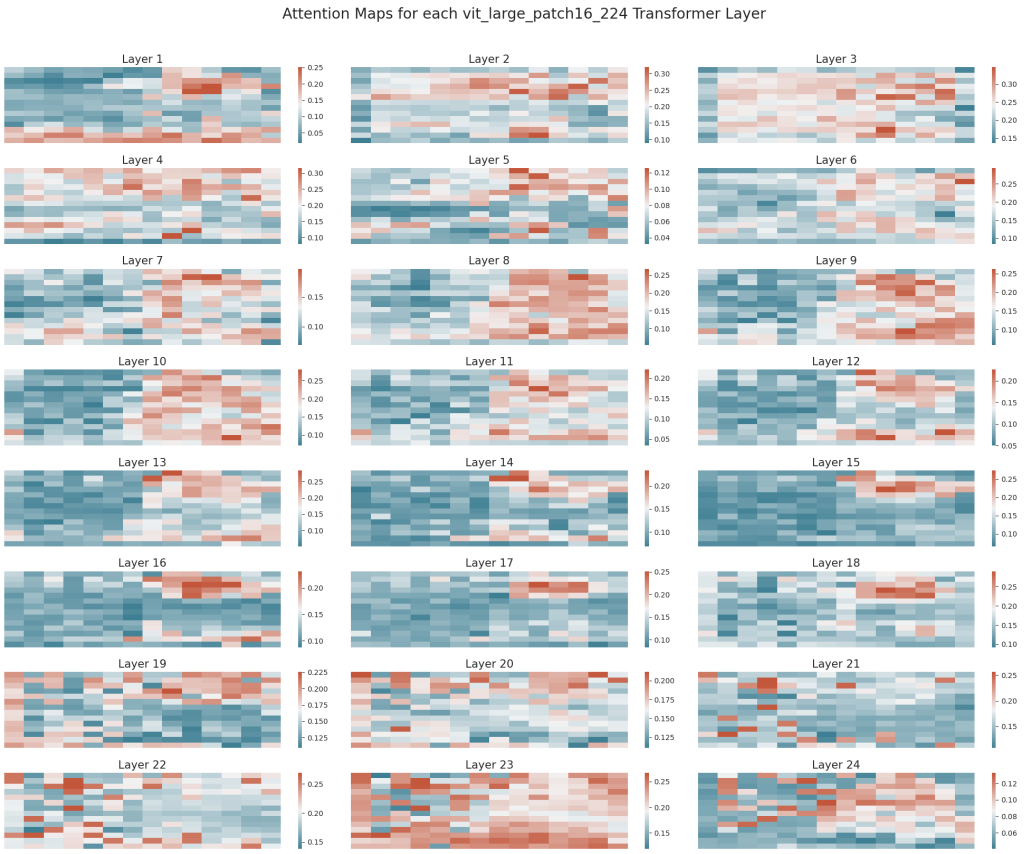
Using the same set of five input images as before, it was observed that the *deit_tiny_patch16_224* model once again stood out in terms of performance. However, the improved precision in attention maps was not limited to this model; all models showed enhanced interpretability with the Gradient Attention Rollout technique.
This technique's effectiveness can be attributed to its ability to suppress noise in the image. Rather than distributing attention across the entire image, the method hones in on specific parts of the image. This targeted approach allows for sharper and more meaningful attention distribution, enabling more efficient image identification. The improved interpretability offered by the Gradient Attention Rollout technique suggests its potential for enhancing understanding and performance of Vision Transformer models.

Attention Maps for each vit_base_patch16_224 Transformer Layer



Attention Maps for each vit_base_patch32_224 Transformer Layer





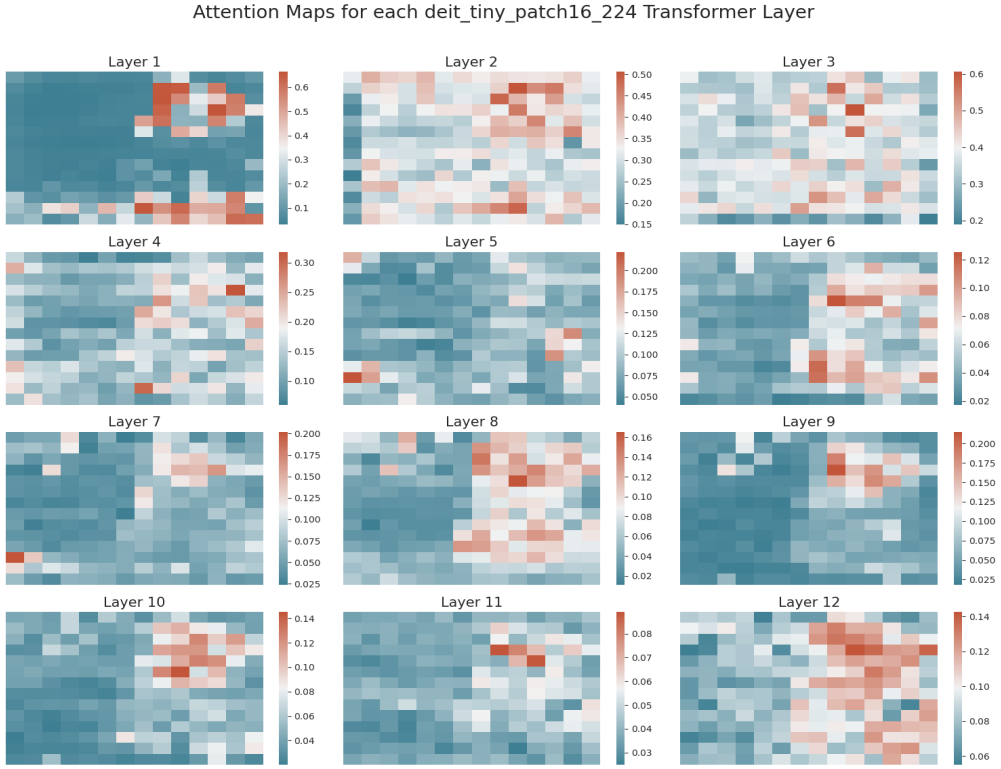


Fig 4.1: Self Attention Map along different layers of ViT Variants

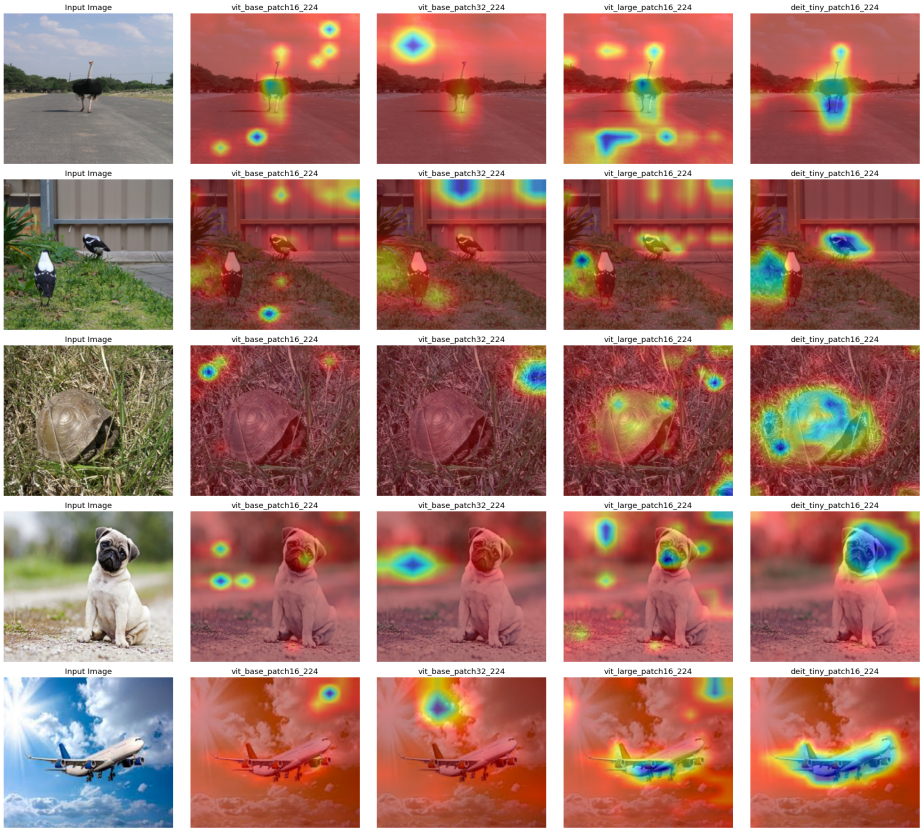


Fig 4.2: Rollout Technique Results

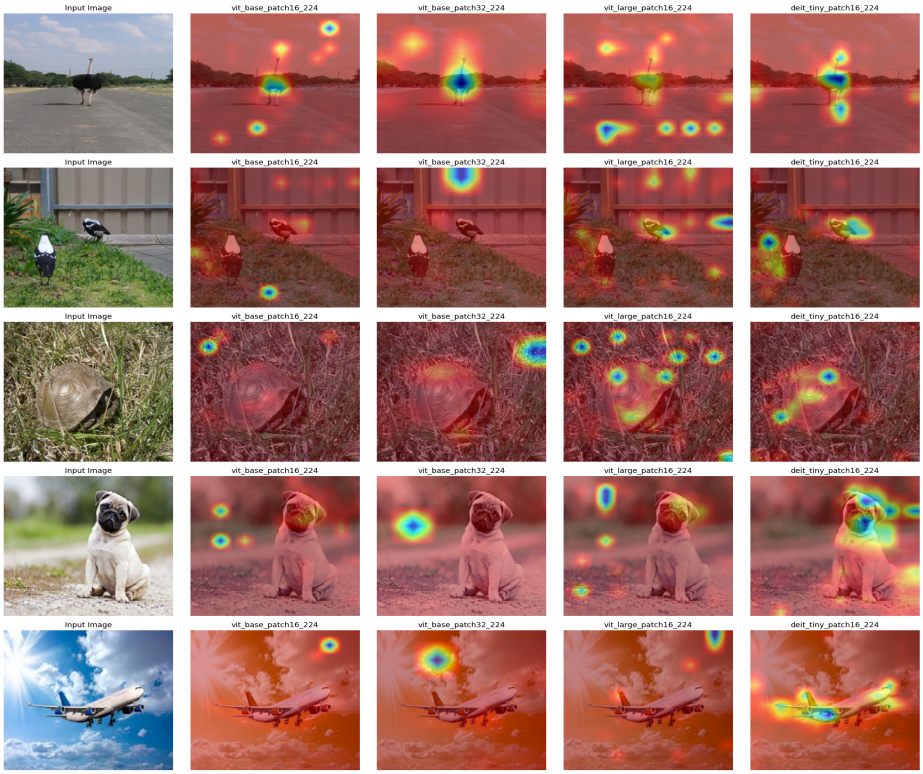


Fig 4.3: Gradient Rollout Technique Results

5. Conclusion and Future Work

Our investigation into Vision Transformer's (ViT) self-attention mechanisms provides valuable insights into its decision-making process. We discovered how the model's interaction with image data at varying transformer layers influences the final output. The Attention Rollout and Gradient Attention Rollout techniques we devised have been pivotal in this process, especially the latter, which demonstrated focused attention distribution and superior interpretability. Among the models explored, DeiT ViT emerged as the best performer.

This work sets the stage for future studies on the interpretability of transformer-based models in computer vision, highlighting the scope for further improvements. Our upcoming endeavours will focus on extending these techniques to other transformer models to enhance our understanding of their data processing.

Furthermore, we aim to fine-tune these methods for scenarios where interpretability is key, like medical imaging or autonomous driving. We're also interested in integrating these techniques into the training process to guide the model towards more interpretable results.

Finally, we'll explore the application of these visualization techniques to other areas, such as object detection, segmentation, or video analysis, with the end goal of making transformer models powerful, transparent, and trustworthy tools for users.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Uszkoreit, J. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[2] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Caron, M. (2020). Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000-6010).

[4] Greydanus, S., Dzamba, M., & Yosinski, J. (2017). Visualizing and Understanding Atari Agents. arXiv preprint arXiv:1711.00138.

[5] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Workshop at International Conference on Learning Representations.

[6] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv preprint arXiv:1708.08296.

[7] Yu, L., et al. (2020). VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers.

[8] Vig, J., et al. (2019). Transformer Interpretability Beyond Attention Visualization.

[9] Jain, S., & Wallace, B. (2019). Attention is not Explanation. North American Chapter of the Association for Computational Linguistics (NAACL).

[10] Vig, J., Belinkov, Y., & Cho, K. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.