Visualising Attention in a Pre-trained Vision Transformer

Gabriel Urbaitis

May 6, 2025

1 Introduction

The objective of this lab is to inspect a pre-trained ViT-B/16 and determine which image regions contribute most to its "dog" prediction. Two inputs are used: one outside on bright grass under a tree, the other inside on a stair against a plain wall. For each image, four heat-maps were produced. First, The starter CLS overlay shows the final-layer, all-heads attention the notebook provided. Second, Layers 1, 6, and 12 are compared side-by-side so one can observe how attention changes with depth. Third, after scanning all twelve layers of each, the three that place the strongest heat on the dog are compared to capture focus at the peak. Finally Abnar & Zuidema's Attention Rollout (2020) is used to trace how information flows through the whole network.

2 Background

2.1 Vision Transformer B/16

Vision Transformers (ViT) take images as input and split them into a sequence of tokens before feeding them through a stack of Transformer Encoder blocks. The base model used here, ViT-B/16, splits a 224x224 RGB image into a 14X14 = 196 non overlapping patches, each one 16x16 pixels. Every patch is flattened, projected to a 768-dimensional embedding, and added to a learned position embedding so the model knows where the patch came from. A single [CLS] (classification) token is inserted at the start of the patch sequence. After 12 Transformer layers the final embedding of that CLS token is treated as a summary of the entire image and passed to a small classifier head.

2.2 ViT Layers

Inside each Transformer layer, the model uses multi-head self-attention: 12 parallel attention heads process the input, each learning to focus on different aspects of the image. One head might emphasize edges, another focus on color, or focus on specific regions like the center. Each patch token, including the [CLS] token, can attend to all other tokens, combining their features using attention weights. This output is then passed through a feed-forward network that processes each token to help the model better capture useful features. Stacking 12 of these layers helps the model learn more complex features step by step: early layers focus on smaller details, while later layers understand the image as a whole.

2.3 Attention Rollout (Abnar & Zuidema 2020)

The raw attention map in the last layer only shows where the CLS is looking at the end, but it ignores the fact that patch features already contain information gathered in earlier layers. Abnar & Zuidema's Attention Rollout [1] helps trace how information flows through the whole network by: First, averaging all attention heads in each layer into one matrix $A^{(l)}$ (size $N \times N$). Second, adding part of the identity matrix to include residual connections: $\tilde{A}^{(l)} = 0.5 \cdot A^{(l)} + 0.5 \cdot I$. Third, Normalizing each row so the values sum to 1. Last, recursively multiplying the matrices layer by layer from bottom to top, to get the final rollout matrix: $R^{(L)} = \tilde{A}^{(L)} \cdot \tilde{A}^{(L-1)} \cdots \tilde{A}^{(1)}$.

3 Experiment

3.1 Data

(a) Ziggy Outside

(b) Ziggy Inside

Figure 1: Ziggy in Outdoor and Indoor Settings

The data samples were two JPEGs of the researcher's late dog, *Ziggy*: one outdoor shot on a grassy lawn, the other an indoor photo taken above a carpeted stair. Each image was resized to 224 × 224 pixels and normalized.

3.2 Methods

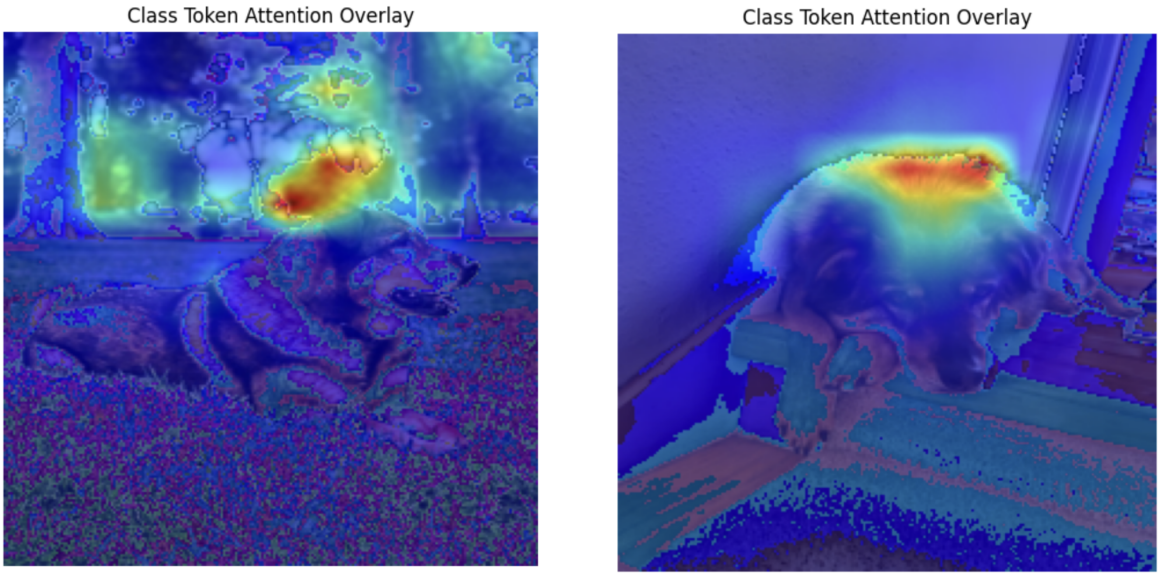
As a baseline, the starter code was used, which averaged the final-layer attention across its 12 heads, extracted the CLS-to-patch row, and overlaid the resulting heat map on the image. To visualize the progression of attention as depth increases, a helper function (`compare_layers_overlay`) was written to display the CLS overlays for Layers 1, 6, and 12 side by side.

Examining all twelve single-layer CLS overlays for each scene, three layers per sample were identified in which the heat map placed its brightest region directly on Ziggy's muzzle or torso. These layers were selected and saved for a sharpened focus comparison.

Finally, Abnar & Zuidema's (2020) Attention Rollout was implemented: attention heads were averaged, a half-scaled identity matrix was added to account for residual connections, rows were re-normalized, and the matrices were multiplied from bottom to top. The CLS row of the resulting rollout matrix was overlaid on each photo using straightforward min-max scaling ($\alpha = 0.6$).

4 Results

4.1 CLS Token Overlay



(a) Outside CLS Token Overlay

(b) Inside CLS Token Overlay

Figure 2: CLS Token Attention Overlays

For the outdoor scene, there are hotspots above Ziggy's back and neck region. Surprisingly, grass and sky patches receive almost equal weight. For the indoor scene, the hottest region sits on the fur above Ziggy's head. Apparently, this is the most dog-specific feature, perhaps because the background textures are similar.

4.2 Layer Progression

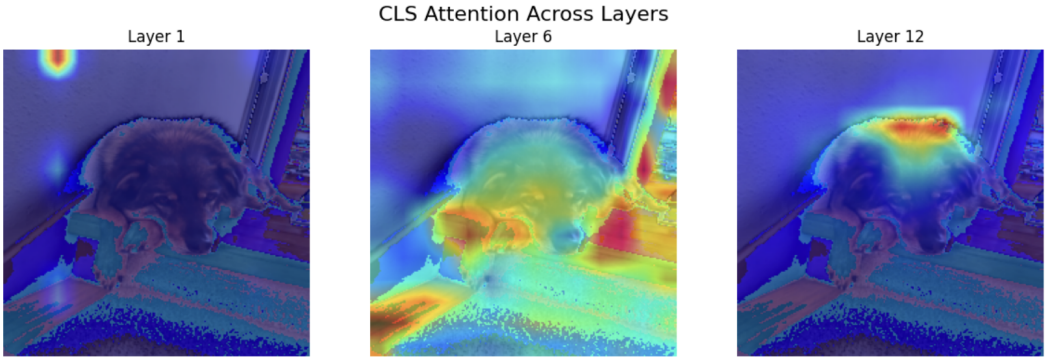


Figure 3: CLS Attention Across Layers – Inside Scene

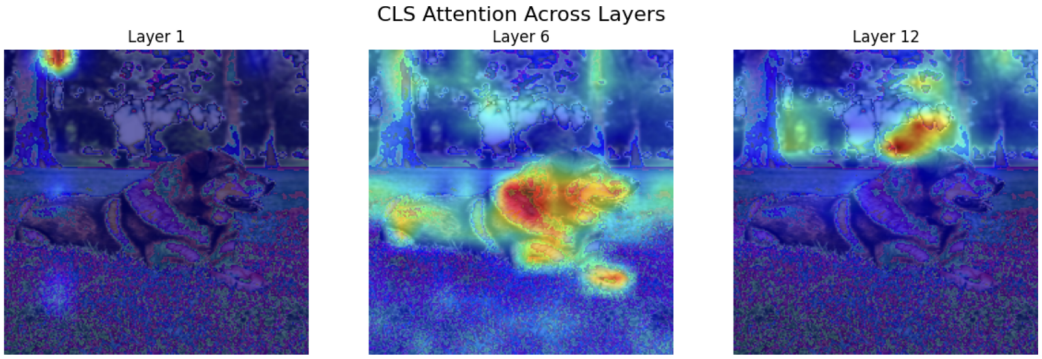


Figure 4: CLS Attention Across Layers – Outside Scene

The hottest patch for the first layer is a random lighting spot on the wall on the indoor scene and a bright spot in the sky for the outdoor scene. This seems to indicate that the heads focus on the highest contrast area before they know the object. In Layer 6, the heat is on the torso outline and paw edges for the indoor scene and the full torso outline for the outdoor scene. The transformer seems to converge on body shape once context is integrated. In Layer 12, the highest head is on an area above and slightly overlapping with Ziggy's head, suggesting the CLS copies from both an object specific patch, and a context patch. Overall, the ViT doesn't just focus on the dog and discard the background, it uses the context from ImageNet to focus on the scene as well.

4.3 Most-Focused Layers

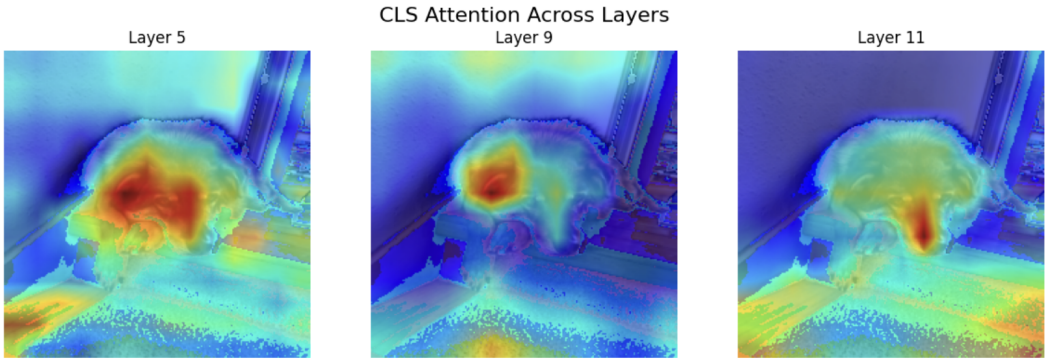


Figure 5: Most-Focused Layers – Inside Scene

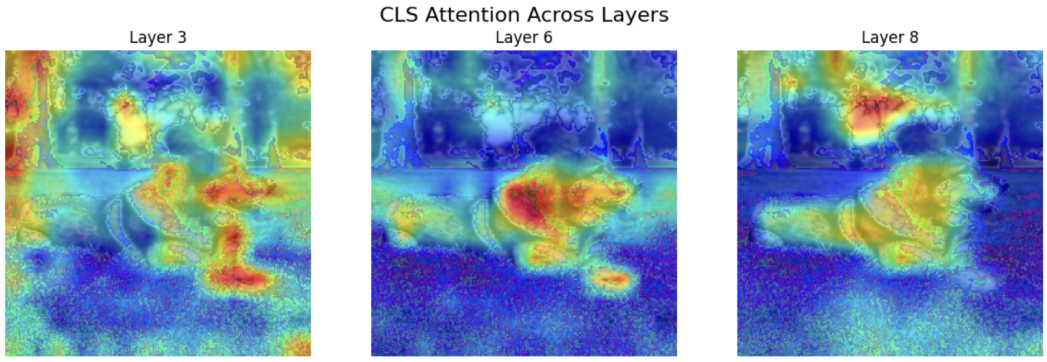
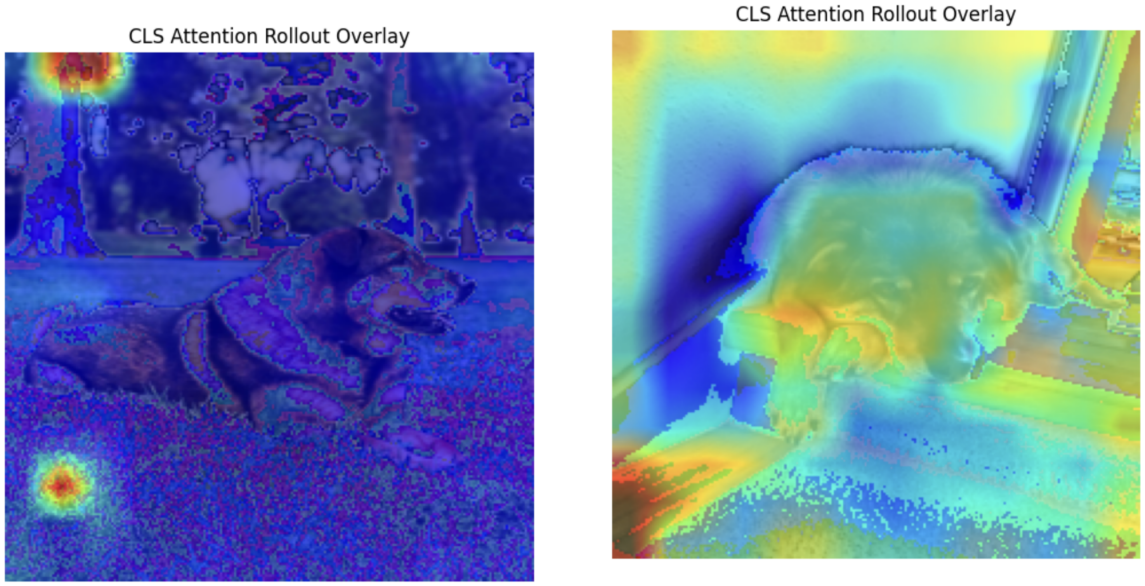


Figure 6: Most-Focused Layers – Outside Scene

For the indoor scene, Layer 5 has heat focused on the head, shoulders and some back fur, with a little focus on the step. By layer 9 the hotspot focuses on the right ear (left in the image), and the wall and carpet are more faded. Layer 11 focuses more, isolating the snout up to the nose tip, and the focus on the step returns. For the outdoor scene, Layer 3 has many focuses, not only the torso, rear leg, but also the trees and a bit of grass. By Layer 6, the heat is mostly distributed across the entire body, with extreme focus on the shoulders and face, but there are still some leaves glowing in green. Layer 8 has an even color across the whole body, and then the most heat on the sky, possibly because the combination of the dog and sky indicates the scene is outside.

Overall, indoors the focus narrows from the body to the ear to the nose, while outdoors the focus narrows on the body. Both re-expand to a context blob, compromising between the dog and the environment.

4.4 Rollout Overlay



(a) Outside Attention Rollout

(b) Inside Attention Rollout

Figure 7: Attention Rollout Comparison

In the rollout overlays we see how information from all 12 transformer layers ultimately flows into the CLS token, and the picture is very different for the two scenes.

For the outdoor scene, the strongest cumulative focus is drawn from two high contrast background patches: the bright hole in the tree, and a sunlit patch of grass in the lower left corner. Ziggy is identified but barely, only a faint outline of his back and fore-leg is above dark blue. This shows that, over depth, the network keeps relying on scene context features, sky and grass are features that frequently show up with dogs in the ImageNet training set, so they still feed the final decision.

For the indoor scene, the Rollout heat is almost the opposite. Ziggy's entire body is yellow-green, and his head and paws are clearly warm spots. Background surfaces like the wall, carpet, and wooden step do contribute, but they top out at fainter colors. With fewer high contrast distractions, the model's long-range information flow concentrates on Ziggy.

Taken together, the two rollouts showcase ViT's strategy: it combines object focus with whatever contextual patches most strongly predict the class. When the background is brightly lit (outdoor), the context patches outweigh the object; when the scene is plainer (indoor), the object outweighs the context.

5 Conclusion

The heat maps tell a consistent story. In both scenes, Layer 1 brightens the single patch with the strongest sky or wall, showing the model starts by scanning for contrast. By Layer 6, the hot zone shifts onto Ziggy's whole body, meaning the patch-to-patch attention has built a foreground outline. Layer 12 narrows further: indoors, the spotlight sits on the snout, outdoors it splits between snout and a bright sky patch, so the model still keeps a background feature when the scene is busy. The rollout map confirms that mix. Indoors, Ziggy dominates the cumulative focus; outdoors two context blobs, sky and grass, end up outweighing Ziggy. ViT-B/16 therefore blends object and scene. Mid-depth CLS overlays give the clearest object outline, though rollout is best when one wants to measure how much context the network relies on.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. URL https://arxiv.org/abs/2005.00928.