

Supplementary Materials: Peeling Back the Layers: Interpreting the Storytelling of ViT

Anonymous Authors

A ADDITIONAL RESULT ON LOWER LAYERS

In this supplementary material, we present the linguistic outputs when averaging layers 0-10 and confusing layers 11-20 in the lower levels of the model. These outputs are displayed in Figures 1 and 2, respectively. This visualization helps to illustrate the differences in linguistic expression resulting from operations in these specific layer ranges.

- It's a photo of a flower that doesn't have a flower in it.
- it was taken with a digital camera and not a point-and-shoot.
- The photo depicts a strange scene of smoke rising from a black background.
- it is not a picture of a person, but a picture of a woman holding a baby in one hand and a man holding a camera in the other hand.
- It's a black and white photo with no people in it, which makes it look like a black and white photo of a person.

Figure 1: The examples of descriptions generated from averaging layers 0-10.

We select five random examples of output from layers 0-10 and 11-20 of the model. It is observed that averaging the outputs from layers 0-10 often results in descriptions that contain contradictions and display a type of disorder. This is likely a consequence of the averaged vector encapsulating a mix of diverse features, which leads to seemingly contradictory statements. For example, a description might simultaneously claim that a scene contains "a flower" and "does not have a flower", reflecting the confused signals derived from the averaged inputs.

In contrast, outputs from layers 11-20 tend to interpret more complex scenarios, featuring overlays of various images. These descriptions often blend black and white with color, supporting the hypothesis that layers 11-20 are more involved in recognizing specific shapes and colors. This behavior indicates that these layers manage more refined visual information, integrating complex features into coherent, whole-image representations.

These observations underscore the distinct roles played by the lower and middle layers of the ViT model in processing visual information. The initial layers handle basic and often contradictory elements, while the subsequent layers organize these into more complex, color-coordinated visual narratives. This delineation of layer functionality enhances our understanding of the hierarchical nature of visual processing within the ViT architecture.

B ADDITIONAL HEATMAP AND LINGUISTIC COMPARISON

In this section, we further explore the interaction between heatmaps and linguistic descriptions, as illustrated in Figure 3. Each layer of

- This photo is a black and white image with a circular pattern in the middle. It looks like an abstract painting, but it's actually a photograph taken by a professional photographer. The background of the photo is dark and blurry, which creates a striking contrast with the rest of the picture. What's strange about this photo is that it appears to have been taken with a digital camera rather than a film camera, which could be a result of poor lighting conditions.
- This is a photo of a black and white swirling pattern. It appears to be a digitally created image, but it's actually a photograph taken with a cell phone camera. The shapes in the background are not perfectly aligned, creating a triangular pattern that looks like a starry night sky. What's strange about this photo is that there is no way to tell what it is or where it came from. You can only guess at what it might be based on its appearance alone.
- It's a black and white photo of a rainbow, but it's actually made up of many different shades of red, green, blue, yellow, orange, purple, and even black.
- The image is composed of different colors and shapes, creating a surreal effect that makes it look like an abstract painting.
- This photo features a complex pattern of swirling shapes and colors that creates a surreal visual effect. It's reminiscent of an abstract painting, which is why it's often referred to as a "fractal" or "fractal art."

Figure 2: The examples of descriptions generated from confusing layers 11-20.

the heatmap highlights distinct focal points: starting with a broad focus on the background, the attention progressively narrows to key areas within the image, and eventually, in the final layers, it tends to cover the entire area uniformly.

Simultaneously, there is a notable transformation in the semantic content of descriptions. Initial layers provide a general overview of the image, which progressively evolves into more specific and detailed descriptions. The correlation between the heatmaps and the linguistic output demonstrates a clear, tandem progression. Initially, the broad heatmap focus underscores a comprehensive grasp of the scene's context. As the layers progress, this focus becomes more targeted towards salient features, mirrored by increasingly precise and detailed linguistic descriptions. In the latter layers, the more uniform heatmap coverage aligns with a plateau in the semantic refinement of descriptions. This uniform attention across the entire image indicates a thorough integration of visual data, leading to stabilized enhancements in the linguistic output. Such enhancements plateau, explaining why substantial improvements in metrics such as CIDEr are not seen beyond certain layers. This dynamic underscores the model's ability to integrate and refine visual and textual information through its multi-layered architecture,

culminating in a nuanced and balanced understanding in the upper layers.

C ABSTRACT TO CONCRETE IN LAYERS 24-28

In this section, we present additional examples of the transition from abstract to concrete descriptions. These subtle differences are primarily identified through GPT-4V scoring, which pinpoints layers 24 to 28 as critical in this transformation. The specifics are illustrated in Figure 4. This detailed representation highlights how nuances in semantic changes are captured and quantified through scoring.

D ADDITION EQUATIONS ON SPECIFIC ATTENTION HEAD

In this section, we explore isolating the contribution of specific attention heads within a layer using a masking technique. This approach modifies the Query (Q), Key (K), and Value (V) matrices to retain only the effects of the selected head, while other heads are effectively ignored by applying a mask. Specifically, a mask matrix M_h is defined where M_h is 1 for parts corresponding to the specific head h and 0 elsewhere. This modification allows for the computation of attention scores and subsequent values to consider only a particular head.

The original formula:

$$\left[\text{MSA}^l \left(Z^{l-1} \right) \right]_j = \sum_{h=1}^H \sum_{i=0}^N x_{ij}^{l,h}, \quad x_{ij}^{l,h} = \alpha_{ij}^{l,h} \left(z_i^{l-1} W_V^{l,h} \right)$$

is modified to include only a specific head h :

$$\alpha_{ij}^{l,h} = \text{softmax} \left(\frac{Q_h K_h^T}{\sqrt{d_k}} \odot M_h \right)_{ij}$$

The formula incorporating a scaling factor is:

$$\left[\text{MSA}^l \left(Z^{l-1} \right) \right]_j = \frac{\|Z^{l-1}\|_2}{\|z_i^{l-1} W_V^{l,h}\|_2} \sum_{h=1}^H \sum_{i=0}^N \alpha_{ij}^{l,h} \left(z_i^{l-1} W_V^{l,h} \right)$$

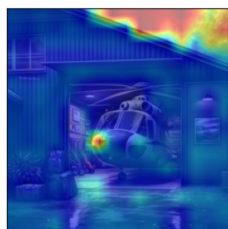
Here, M_h is the mask matrix, which is 1 only for the selected head h' and 0 (or a very small value such as $-\infty$) elsewhere. This ensures that after the softmax, $\alpha_{ij}^{l,h}$ values for non- h' heads are very small (near zero), effectively focusing the output's contribution solely from head h' . $\|Z^{l-1}\|_2$ represents the L2 norm of the combined output from all heads prior to applying the mask. $\|z_i^{l-1} W_V^{l,h}\|_2$ is the L2 norm of the output from the selected head h , scaled by the Value matrix weights. $\alpha_{ij}^{l,h}$ is adjusted by the mask matrix M_h , ensuring only contributions from the specific head h are considered.

By introducing a scaling factor, we can accurately reflect the contribution of a single attention head while considering the aggregation from all heads. The L2 norm acts as a scaling factor to normalize the contributions from the selected head, ensuring its influence is appropriately adjusted relative to the original vector magnitude from all heads. This method not only enhances the precision of our analysis regarding the impact of individual attention heads but also facilitates future combinations of multiple heads by simply adjusting the mask matrix.

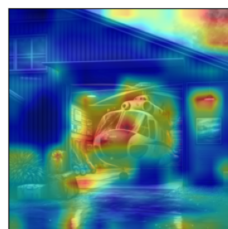
We also present additional heatmaps of decomposed attention heads in this section, as illustrated in Figure 5. It is evident that each attention head has its specific focus areas. By identifying these differences, we can prune those heads we deem to have minimal impact on subsequent layers. This approach allows us to streamline the model by retaining only the most effective attention mechanisms, thus enhancing the overall efficiency and interpretability of the model. This targeted pruning strategy not only optimizes the model's performance but also sheds light on the functional importance of individual attention heads within the architecture.



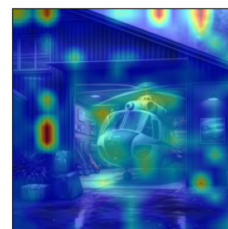
Original image.



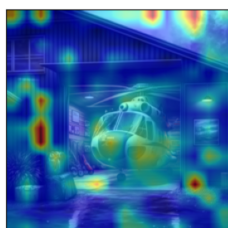
Layer0: It's a picture of a man in a bathing suit and a woman in a bikini.



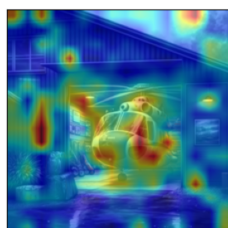
Layer4: It was taken by a camera with no flash.



Layer20: It's not a picture of a car, it's a picture of a train station.



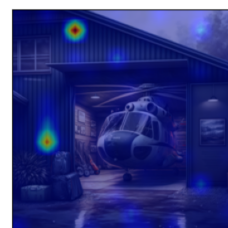
Layer24: The image depicts a strange scene with an automobile in the middle of a parking lot on a cloudy day. The car is parked in front of a building with a large, futuristic-looking skyscraper in the background. However, it's hard to tell what kind of vehicle it is from the photo itself.



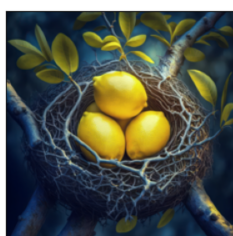
Layer27: The photo depicts a strange scene with an airplane flying in front of a green building. It looks like a spacecraft, but it's actually a car that has crashed into the building and is stuck in the garage.



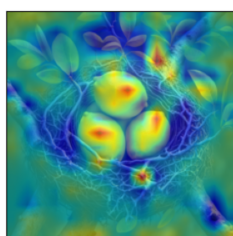
Layer30: The photo shows a plane inside a garage, but it's not an airplane. Instead, it's a helicopter, which makes it look more like a spaceship rather than a plane. It's also hard to tell what kind of plane it is, since it doesn't have a tail or wings. However, it appears to be a military aircraft with a large propeller attached to the end of its fuselage.



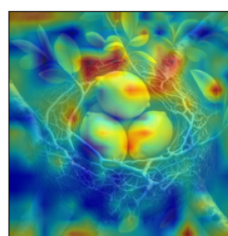
Layer38: The photo shows a helicopter in a garage, but it doesn't look like a helicopter at all. Instead, it appears to be a painting of a hangar with a helicopter inside. What's strange about the photo is that there are no other people in the picture, so we can't really tell if this is a real image or a fake one.



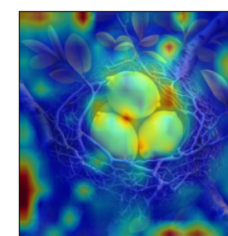
Original image.



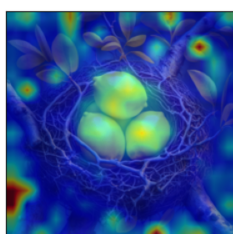
Layer0: It's a black and white photo of a woman in a bathing suit.



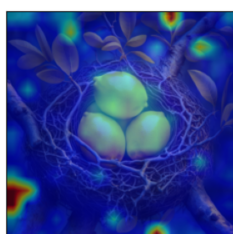
Layer4: It's a photo of a man and a woman in front of a tree.



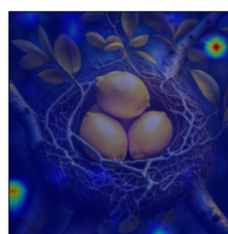
Layer20: It's a photo of a bunch of bananas sitting on top of a piece of luggage in a train station.



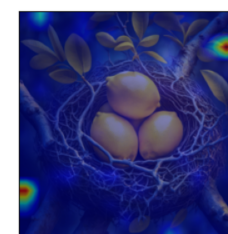
Layer24: The photo shows a group of oranges nestled in a tree trunk. However, what's strange about this photo is that the oranges appear to have fallen out of the tree and are still attached to the trunk. This may be due to the fact that these oranges were left inside the tree for a long period of time, making them difficult to remove from the tree.



Layer27: The image of the yellow peaches nestled in a tree branch is a bit odd, as it appears to be a natural arrangement that has been created by nature. However, there are some strange details about the photo, including the fact that the peaches are nestled in a tree branch, rather than on a tree trunk.



Layer30: The image depicts a group of oranges nestled in a tree trunk, creating a surreal scene that captures the essence of nature's natural beauty. However, there is something strange about this photo, as the oranges are nestled inside a tree trunk, rather than in a tree or on a tree branch.



Layer38: This photo features three lemons nestled in a bird's nest, creating a surreal scene that captures the beauty of nature. The image was taken from a digital camera, and it has been digitally enhanced to create a more realistic effect.



Figure 3: The original image, heatmaps, and semantic descriptions for each layer reveal noticeable shifts in focus areas and a gradual transition in semantic expressions from abstract to concrete.

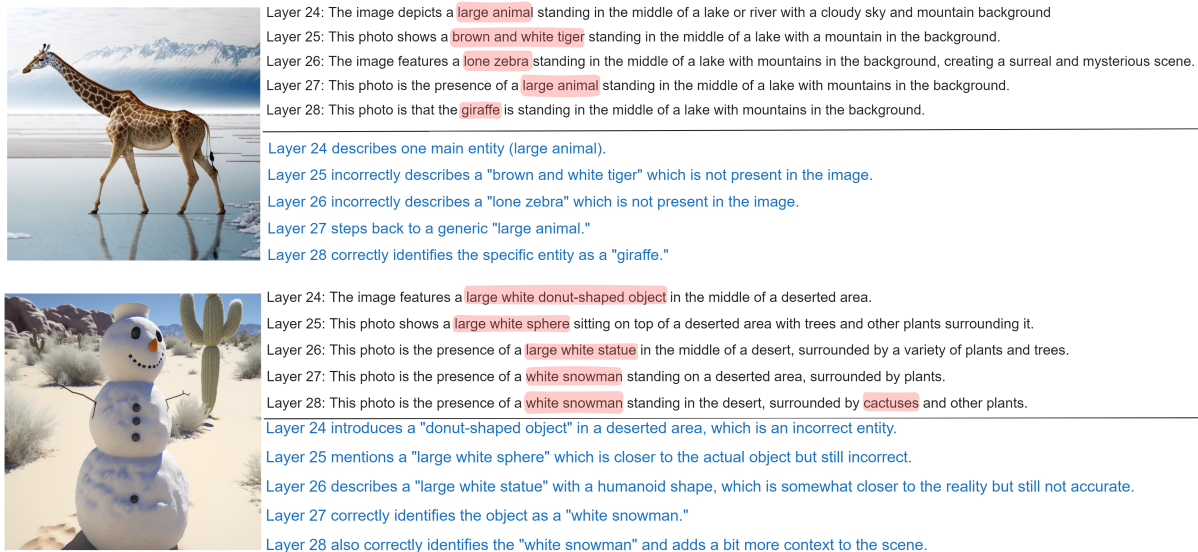


Figure 4: The examples below illustrate the transition from abstract to concrete descriptions, where the black text represents the original description and the blue text provides explanations aligned with the phenomenon of transitioning from abstract to concrete.

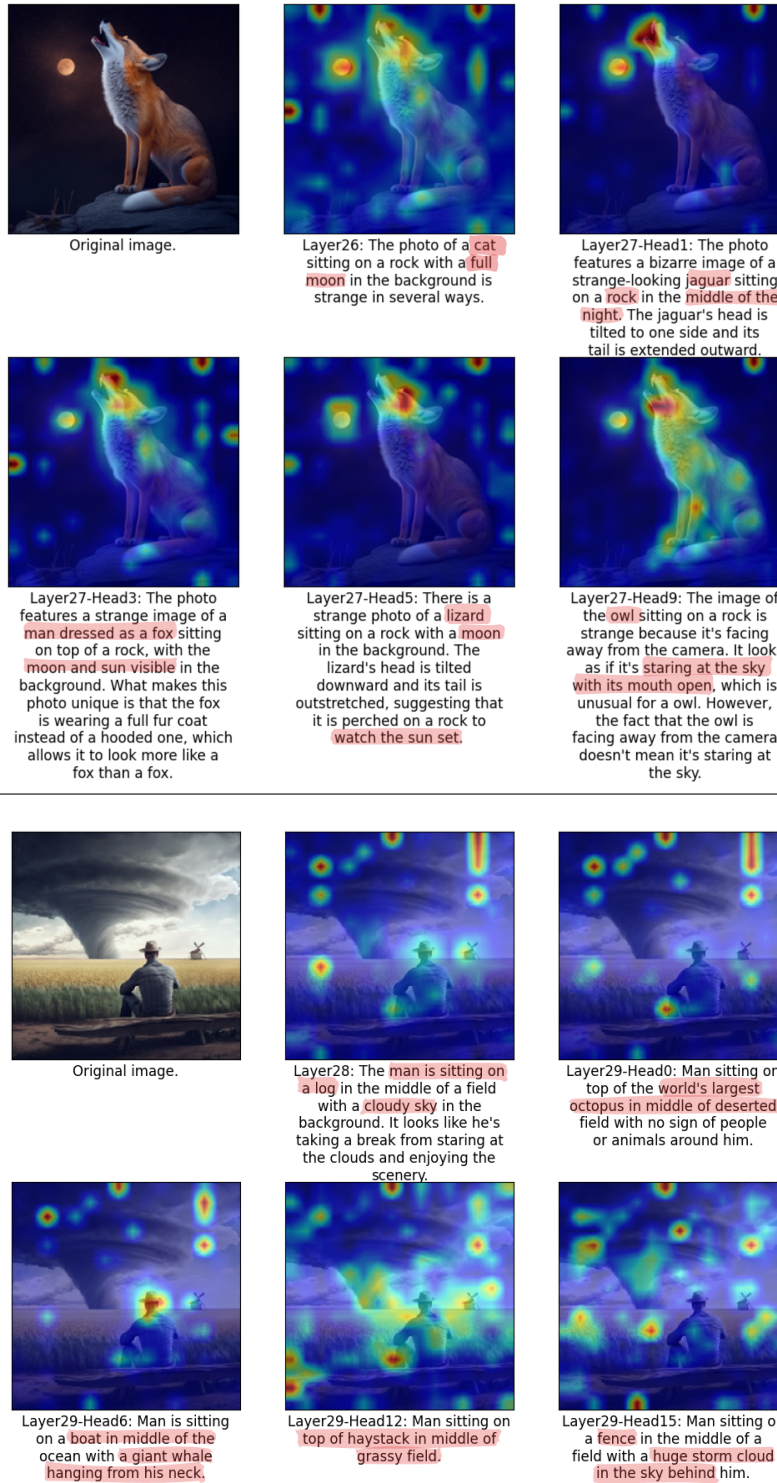


Figure 5: The original image, outputs from the previous layer, and attention heads from the subsequent layer exhibit a correlation between heatmaps and textual descriptions that is specifically linked to specific heads.