Analyzing Vision Transformer Patch Attention and Robustness

Khadeeja Toseef

8th September 2025

GitHub Link : Khadeeja's GitHub

Introduction

Vision Transformers (ViT) have emerged as powerful models for image classification, processing images by splitting them into patches and applying self-attention across all patches. In this work, we use a pretrained ViT-base model to study how patch-level attention contributes to classification and how the model handles missing patch information. We focus on visualizing the patch attention, masking patches at inference time (both randomly and in a structured way), and comparing two pooling methods for classification. Our goal is to understand the robustness of ViT to occlusions and the role of the [CLS] token versus average pooling in producing image embeddings.

Methodology

We employ the pretrained `google/vit-base-patch16-224` model, which is a ViT with 12 layers and 16×16 patch size, fine-tuned on ImageNet. We process a test image of a cat by first applying a ViT image processor (resizing and normalization) and then passing the image through the model with attentions enabled. The model outputs classification logits and the attention tensors for each layer.

Attention Rollout Visualization

To interpret the ViT's focus, we compute a cumulative attention map using the "attention rollout" technique. Specifically, for each transformer layer we average attention over heads and add the identity matrix (to include residual self-attention). We then normalize each layer's matrix row-wise and multiply these matrices across all layers. This yields a final attention matrix representing the influence of each input patch on the [CLS] token. Extracting the row corresponding to the [CLS] token and reshaping it into a square grid produces a patch-wise attention map. We normalize this map to [0,1], resize it to the image dimensions, and apply a Gaussian blur to generate a heatmap overlay. This overlay highlights regions of the image that most strongly influence the classification.

Patch Masking Experiments

To test the robustness of the ViT to missing information, we mask a fraction of the input patches before classification. We implement two masking strategies: (1) *Random masking*: we randomly select 30% of the patch tokens to mask out; (2) *Center masking*: we mask out a contiguous square region of patches centered in the image, covering 30% of patches. In both cases, we replace the selected patch token embeddings with zeros. We then prepend the [CLS] token embedding, concatenate with the (masked) patch tokens, and feed this sequence through the transformer encoder and classifier to obtain a prediction. We compare the predicted class label under each masking strategy to the original unmasked prediction.

Pooling Comparison

We also compare two methods of pooling patch features into an image-level representation for classification: (a) using the [CLS] token's hidden state, and (b) taking the mean of all patch token hidden states. For each method, we pass the resulting pooled vector through the model's linear classifier to predict the class. This allows us to assess how different pooling choices affect the final classification.

Results

For the chosen test image of a cat, the ViT model produced a top-1 classification label corresponding to the cat (e.g., "tabby cat"). The attention rollout overlay (not shown) highlights the cat's body region, indicating that the model's global self-attention focuses on the main object when making its decision. In the patch masking experiments, we found that the original (unmasked) image and the two masked versions all yielded the same predicted label. Specifically, the original prediction was the correct cat class, and both random masking and center masking with 30% of patches removed did not change the top-1 prediction. This suggests that, for this example, the ViT was robust to the removal of a substantial fraction of input patches. In particular, random masking (uniformly distributed erasures) had minimal impact because the model could compensate using the remaining context. Center masking also did not alter the prediction in this case, although removing the central region (where the object is often located) is expected to be more challenging. For pooling, both the [CLS] token and mean-pooling of patch embeddings yielded the correct class label for the image. There was no difference in prediction accuracy for this example, indicating that either method can suffice for standard image classification when the model is already trained for this task.

Discussion

Our experiments indicate that the ViT's global self-attention mechanism lends robustness to random occlusions: because each output token attends to all patches, missing information in some patches can often be inferred from others. In contrast, structured occlusion (such as masking the central region) tends to target the most informative patches. In our experiment the ViT still correctly classified the image under central masking, but in general we expect structured occlusions to have a stronger effect on accuracy if key object features are removed. The attention overlay confirms that the model heavily attends to the object; thus, removing those patches can degrade performance more than removing peripheral patches. Regarding pooling strategies, the [CLS] token is designed during pretraining to capture global information for classification, and indeed it is directly optimized for downstream tasks. Mean pooling aggregates information from all patches, which can provide a more comprehensive representation of the image but may dilute specific discriminative features. In practice, we observed that both methods worked for this example. However, one might choose [CLS] pooling when fine-tuning for classification (as it is more efficient and aligned with the pretraining objective) and consider mean pooling when broader image representation is needed (e.g., for tasks relying on general similarity or retrieval).

Conclusion

We have analyzed how a pretrained Vision Transformer processes a test image through its attention mechanism and classification layers. The visualization of patch attention confirmed that the model attends to salient object regions. The model proved relatively robust to missing patches: masking 30% of patches (even in the center) did not change the classification for our example. Finally, we showed that both [CLS] token pooling and mean-pooling of patch features can recover the correct label, though their use cases may differ. This study provides insight into the internal workings of ViT and its resilience to partial occlusion, as well as considerations for how to extract image embeddings for downstream tasks.

Figure 1: Attention Based rollout overlay.