Disentangling Visual Transformers: Patch-level Interpretability for Image Classification

Supplementary Material

1. Model architecture details

To compare our architecture as fairly as possible with standard methods (*i.e.*, DeiT [8]), we follow the same hyperparameter selection for both the base and small versions as their ViT equivalents. That is, HiT base (HiT-B) follows the same architecture as ViT-B. Similarly, HiT small (HiT-S) follows the same hyperparameters as ViT-S. The only differences are the pooling layers, the initial patch size, and that we have removed the last MLP block as it is not used. Tab. 1 shows the architecture hyperparameter choices compared to Visual Transformers.

Version	Layers	Feature dimension	Heads	Pooling layer location	Patch dimension	Parameter count
HiT-B	12	768	12	$[4, 8]$	8	81.8
ViT-B	12	768	12	-	16	86.9
HiT-S	12	384	6	$[4, 8]$	8	20.8
ViT-S	12	384	6	-	16	22.1

Table 1. HiT base and small hyperparameter configurations

2. Datasets

As for the dataset, we evaluated HiT on six diverse image classification datasets: i) ImageNet [2]: A large-scale dataset with 1.2 million images and 1,000 classes, often used as a benchmark. ii) CUB-2011 [9]: A challenging dataset containing 200 bird classes with only 30 training samples per class on average. iii) Stanford Dogs [3]: A dataset with 120 dog classes and 10,000 training and test images. iv) Stanford Cars [4]: A dataset featuring 196 car classes with 8,100 training and validation examples. v) FGVC-Aircraft [5]: A dataset of 100 airplance classes with 10,000 images. vi) Oxford-IIIT Pets [6]: a 37 category dataset with roughly 200 images per class.

3. Insertion-Deletion Curves

In Fig. 1, we present the curves from the post-hoc comparison experiment detailed in § 4.3 of the main document. HiT consistently outperforms both GradCAM [7] and Rollout Matrix [1] across all datasets. Interestigly, GradCAM achieves performance comparable to our method on all datasets except ImageNet [2]. We hypothesize that this correlation stems from GradCAM being computed on the final layer tokens, which our analysis shows are the most important (Fig. 5 in the manuscript), except for ImageNet.

4. Evaluating HiT without Pooling Layers

We conducted experiments similar to those in the main manuscript to analyze the positive and negative impact of removing pooling layers. As demonstrated in § 4.7 of the paper, pooling layers are essential for improving top-1 accuracy performance. However, their inclusion increases the size of the explanations. First, we explore this phenomenon quantitatively in sections 4.1 and § 4.2. Later, we will explore the qualitative differences in § 4.3.

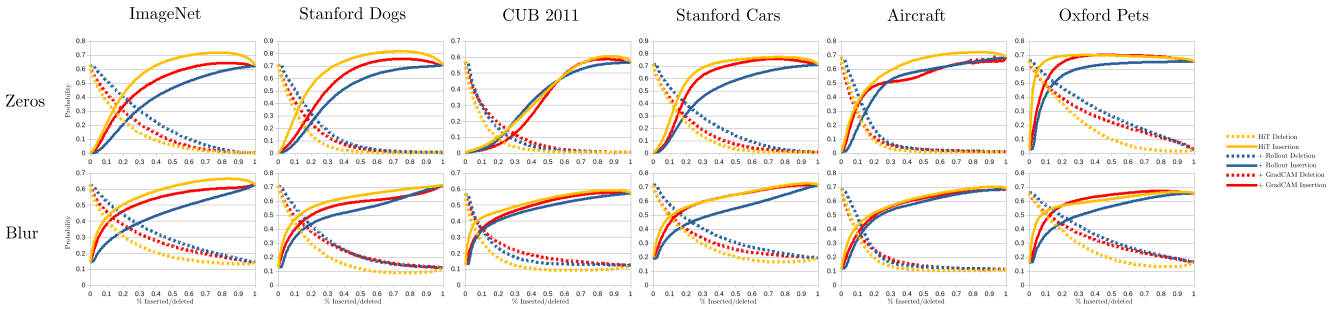


Figure 1. **Comparing HiT and alternative post-hoc methods.** This experiments reflects the interpretable advantages with respect to traditional post-hoc methods.

4.1. Interpretability Trade-off

First, we explore the interpretability gains of HiT without pooling layers. We compare both HiT versions using the normalized insertion-deletion curves on all tested datasets, illustrated in Fig. 2. From a quantitative point of view, HiT without any pooling layer is even more interpretable than our proposed architecture. Interestingly, both curves behave similarly, showing a decrease in the insertion probability curve and an increase in the deletion probability curve during their final steps. This is due to the insertion (or deletion) of tokens that adversely affect the model's prediction.

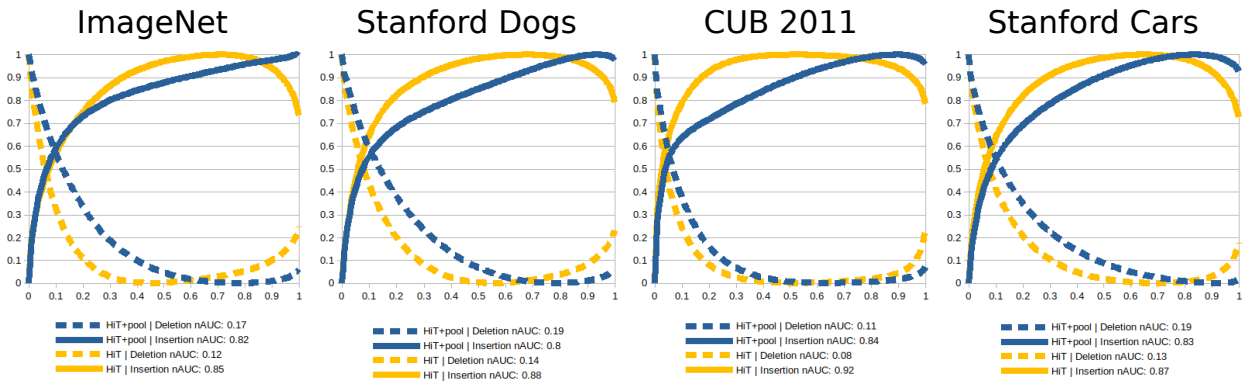


Figure 2. Caption

We also tested GradCAM [7] and the Rollout [1] Matrix directly on our pooling-free HiT architecture, with results shown in Table 2. The interpretability gap between post-hoc methods and HiT saliency maps is more pronounced compared to our standard HiT architecture.

Finally, in Tab. 3, we show the performance on the tested datasets. Without any surprise, the loss in performance is major, making it a less appealing option in contrast to our original architecture when computation power is needed.

4.2. Layer-wise Contributions

Next, we investigate the ability of the pooling-free HiT to analyze layer-wise contributions. Fig. 3a illustrates the layer contribution per dataset, while Fig. 3b plots the ablation results on ImageNet [2]. Similar to HiT with layer pooling, the most significant contributions come from the final layers. However, unlike the HiT version with pooling layers, all trained models appear to weight their final predictions equally across the last three layers.

4.3. Qualitative Comparison

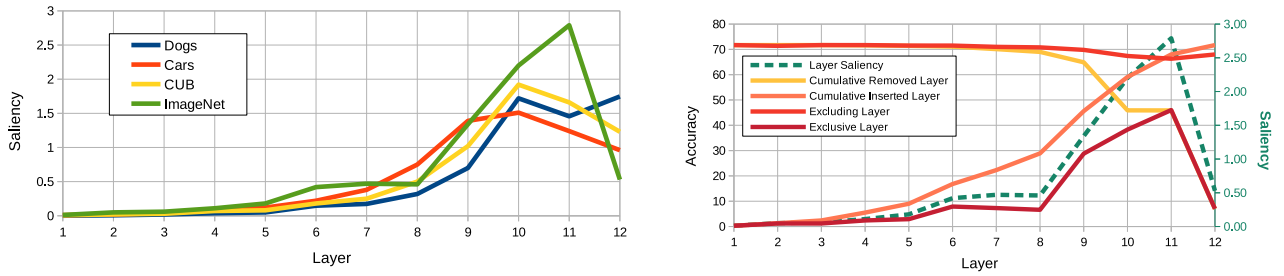
Finally, we qualitatively show the difference between the pool-free HiT and our original version saliency maps in Fig. 4 in ImageNet [2]. As expected, removing the pooling layers produces finer saliency maps.

Method	ImageNet		CUB 2011		Stanford Cars		Stanford Dogs	
	Ins-Z (↑)	Del-Z (↓)	Ins-Z (↑)	Del-Z (↓)	Ins-Z (↑)	Del-Z (↓)	Ins-Z (↑)	Del-Z (↓)
HiT	**0.65**	**0.08**	**0.56**	**0.04**	**0.72**	**0.05**	**0.64**	**0.07**
HiT + Rollout	0.39	0.21	0.43	0.09	0.49	0.12	0.47	0.19
HiT + GradCAM	0.36	0.15	0.40	0.09	0.34	0.11	0.40	0.15
	Ins-B (↑)	Del-B (↓)	Ins-B (↑)	Del-B (↓)	Ins-B (↑)	Del-B (↓)	Ins-B (↑)	Del-B (↓)
HiT	**0.67**	**0.16**	**0.59**	**0.11**	**0.65**	**0.15**	**0.62**	**0.18**
HiT + Rollout	0.47	0.31	0.50	0.22	0.50	0.29	0.50	0.32
HiT + GradCAM	0.48	0.29	0.52	0.21	0.51	0.29	0.52	0.31

Table 2. **HiT and Explainability methods:** We quantitatively compare HiT maps and those created by GradCAM and the modified rollout matrix (mean attention). The assessment shows that HiT maps are in fact more faithful to those generated by GradCAM and the rollout matrix. Higher insertion is better, while lower deletion is better. Ins and Del refers to the Insertion and Deletion metrics, respectively. Z is the zero-corrupted image, while B is the blurred corruption strategy.

Model	Pooling?	ImageNet	CUB	Dogs	Cars
HiT-S	χ	67.3	76.1	77.1	83.9
	✓	71.4	76.1	80.3	85.2
HiT-B	χ	71.5	76.3	80.2	84.7
	✓	75.0	79.0	86.8	86.2

Table 3. **Top1 Accuracy.** Including pooling layers provides a clear advantage in terms of raw performance. However, this performance gain comes at the cost of reduced interpretability.



(a) Layer Saliency.

(b) Empirical validation of layer-wise saliency.

Figure 3. As in the main manuscript, we assess our pooling-free HiT layer contribution. Effectively, HiT can discover the contributions for each layer.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 1, 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2

[3] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 1

[4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[5] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 1

[6] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam:

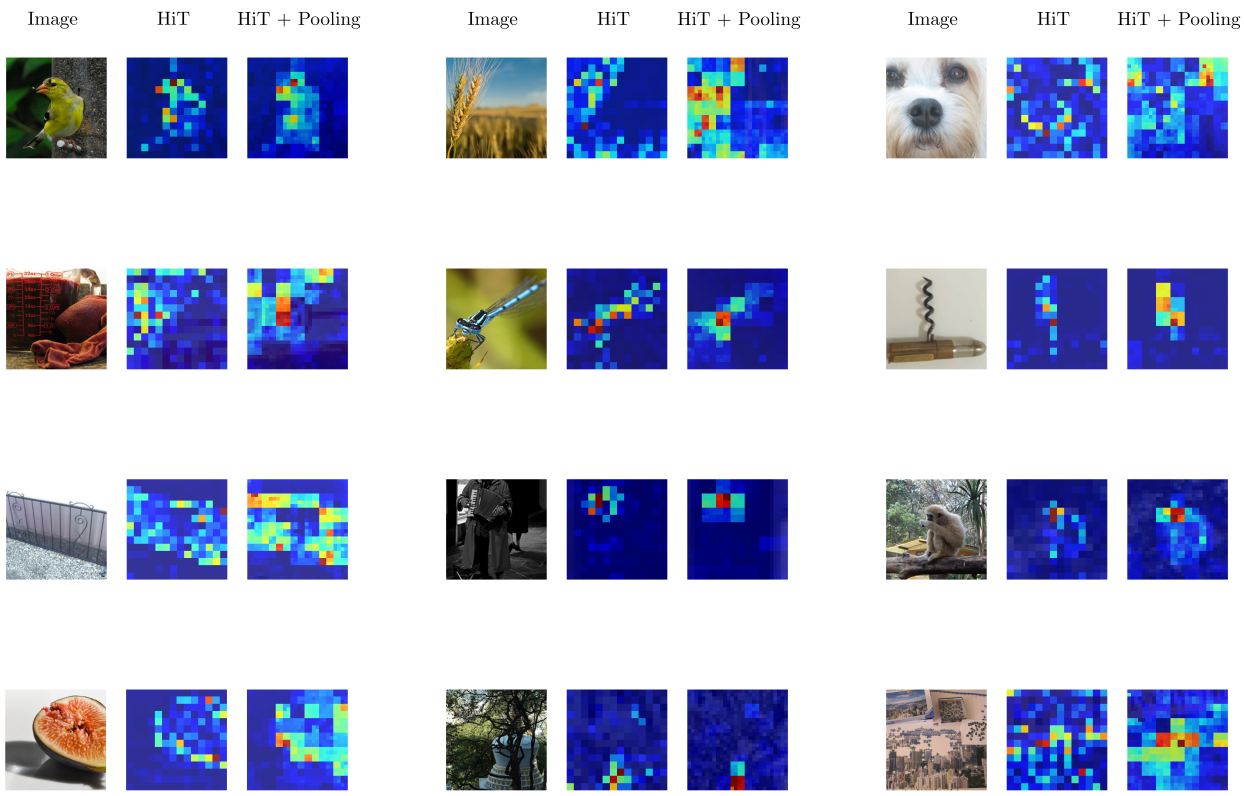


Figure 4. **Qualitative Examples:** we visually show the difference between HiT saliency maps with and without pooling layers on some correctly classified images from the ImageNet dataset.

Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 1, 2

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers; distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1