[Re] Improving Interpretation Faithfulness for Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

This work aims to reproduce the results of Faithful Vision Transformers (FViTs) proposed by Hu et al. (2024) alongside interpretability methods for Vision Transformers from Chefer et al. (2021) and Xu et al. (2022). We investigate claims made by Hu et al. (2024), namely that the usage of Diffusion Denoised Smoothing improves interpretability robustness (1) to attack in a segmentation task and (2) to perturbation in a classification task. We also extend the original study by investigating the authors' claims that adding DDS to any method can improve its robustness under attack. This is tested on baseline interpretability algorithms and the recently proposed Attribution Rollout method. In addition, we measure the computational costs and environmental impact of obtaining an FViT through DDS. Our results agree broadly with the original study's findings, although minor discrepancies were found and discussed.

1 Introduction

By adopting attention mechanisms that constitute transformers (Vaswani et al., 2023), the field of computer vision has experienced significant breakthroughs in tasks such as image recognition (Dosovitskiy et al., 2021), object detection (Zhu et al., 2021), image processing (Chen et al., 2021) and semantic segmentation (Zheng et al., 2021). A notable success is the vision transformer (ViT) proposed by Dosovitskiy et al. (2021) which, in addition to its state-of-the-art performance, provides an attention vector which can be used to explain its decision-making. However, this approach is not robust to slight input perturbations, leading not only to false classification but also to attention vectors that are misleading in interpretability.

Traditional post-hoc interpretability methods have widely been used to analyse ViTs by highlighting important regions in an input image based on attention or gradient-based explanations. They obtain the intermediate self-attention matrices of the model for an input and aggregate them alongside heads and layers. Those aggregations can happen using various components of the model. Raw Attention (Vaswani et al., 2023) directly uses the attention scores from the last self-attention layer to determine importance (Vaswani et al., 2023). GradCAM (Selvaraju et al., 2019) uses gradient information obtained during backpropagation to determine the importance of each pixel (Selvaraju et al., 2019). Rollout (Abnar & Zuidema, 2020) assumes that all attention heads contribute equally and averages them to estimate importance (Abnar & Zuidema, 2020). Layer-wise Relevance Propagation (LRP) propagates relevance scores backwards through the network to estimate pixel importances (Binder et al., 2016). Finally, Transformer Attribution (TA) integrates both gradients and relevance scores to find important pixels (Chefer et al., 2021). The methods output relevance scores for each pixel of the input image, which can be converted into an attention heat map to visualize how relevance regions.

Despite the progress of interpretability methods, the vulnerability to adversarial attacks remains a critical limitation, as small input perturbations can drastically alter their explanations, undermining their reliability. Hu et al. (2024) address this issue by proposing a Faithful ViT (FViT), a vision transformer that is robust to both performance and interpretability concerns. Unlike standard ViTs, where adversarial perturbations significantly affect interpretability, FViTs ensure that explanations remain stable even under such attacks.

To convert a ViT into an FViT, they propose Denoised Diffusion Smoothing (DDS), a novel plug-in method that combines randomized smoothing and diffusion-based denoising. Specifically, it adds Gaussian noise to an input sample, which is then processed through a guided diffusion model to remove the noise. The resulting samples are then fed into a ViT of choice for further analysis.

FViTs also contribute to the ongoing debate between post-hoc and model-based interpretability. While post-hoc methods analyse pre-trained models after training to explain predictions, they often lack robustness to perturbations. In contrast, FViTs integrate interpretability directly into the model, promising that explanations remain stable under adversarial attacks, thus improving both task performance and model faithfulness.

This paper focuses on reproducing Hu et al. (2024)'s experiments, extending the original work by investigating and applying DDS in novel ways, and evaluating the soundness of their main claims.

2 Scope of reproducibility

Hu et al. (2024) claim that (A) the applying DDS to any ViT transforms it into an FViT, and that (B) a FViT, unlike a standard ViT, is robust against adversarial attacks both in task utility and, more importantly, in model interpretability.

To support their claims, the authors provide rigorous mathematical proofs and conduct a series of experiments. Our paper focuses on evaluating the experiments and the main claims derived from them. The following claims are tested:

- **Claim 1:** Applying an interpretability method to an FViT produces more robust explanations than a ViT, particularly under adversarial attacks in an image segmentation task.
- **Claim 2:** Applying DDS to a ViT model, resulting in an FViT, increases the model's classification robustness under input perturbation in an image classification task.

Furthermore, this paper also expands on the original work. Attribution Rollout (AR) (Xu et al., 2022), a new interpretability method, was shown to offer several advantages over TA (Chefer et al., 2021). Since TA was originally used with the DDS (Hu et al., 2024) and because no study examined DDS with AR, this paper aims to fill that gap. Furthermore, Hu et al. (2024) claim that DDS is a plug-in method to improve performance and robustness against adversarial attacks. Therefore, the addition of DDS to the baseline interpretability methods was also investigated. Namely, our extensions are:

- **Extension 1:** Investigating whether DDS with Attribution Rollout produces better interpretability results than DDS with Transformer Attribution in an image segmentation task under adversarial attack.
- **Extension 2:** Investigating whether DDS is a plug-in method for improving the performance and robustness by combining it with Attribution Rollout and baseline interpretability methods in an image segmentation task under adversarial attack.

Lastly, this paper brings to light the possible environmental impacts of applying DDS to ViTs. An investigation of the energy usage of DDS compared to other methods was missing in the original study and was therefore performed. Hence, our study also reports these findings:

- **Impact 1:** FViT's environmental impact is significantly higher than that of all other methods.

3 Background

Faithfulness
Faithfulness refers to how accurately an explanation represents the true reasoning process of a model when making a decision (Jacovi & Goldberg, 2020). It is based on three distinctive assumptions:

- **The Model Assumption:** *Two models will make the same prediction if and only if they use the same reasoning process.*
- **The Prediction Assumption:** *On similar inputs, the model makes similar decisions if and only if its reasoning is similar.*
- **The Linearity Assumption:** *Certain input parts are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent of each other.*

Robustness
ML model robustness denotes the capacity of a model to sustain stable predictive performance in the face of variations and changes in the input data (Braiek & Khomh, 2024). Specifically, this paper focuses on adversarial robustness; the extent to which models resist adversarial perturbations.

4 Methodology

Two experiments were carried out to investigate the two claims: an image segmentation task under a projected gradient descent (PGD) adversarial attack, and an image classification task under positive and negative perturbation, as described by Hu et al. (2024) in addition to Chefer et al. (2021). For both tasks, a pre-trained ViT model was used along with interpretability methods from Chefer et al. (2021) as well as Attribution Rollout by Xu et al. (2022); the ViT model combined with an interpretability method produces attention heat maps, which are used to create image segmentations in the segmentation task as well as inform the regions for the perturbation attacks in the classification task. A PGD attack in the classification task, which was implemented in the original study, was not investigated due to computational constraints. The energy usage reported by the cluster was also measured, which was combined with national publicly available estimates on grid carbon efficiency, to estimate the environmental impact.

Furthermore, the original study implemented a qualitative test of the interpretability methods under an attack; attention heat maps computed by the interpretability methods were provided before and after the attack. We recreated these results following the provided demo implementation.

Moreover, we extend the study by investigating the plug-in claims of DDS as well as the potential improvements of AR. Consequently, AR and the baseline methods were investigated with DDS in the image segmentation task. The classification task was not investigated due to computational constraints.

The code used for the experiments was built on top of Hu et al. (2024) who extended the code from Chefer et al. (2021) by providing the qualitative demo. We further extended Chefer et al. (2021)'s code by adding PGD to the image segmentation task and adding DDS to both tasks based on the demo implementation by Hu et al. (2024). We also added the Attribution Rollout interpretability method to the segmentation task following the implementation by Xu et al. (2022).

4.1 Model descriptions

Following the original study, a ViT-Base model introduced by Dosovitskiy et al. (2021) was used. It was pre-trained on the ImageNet-21k dataset and uses a resolution of 224x224. [1]. Moreover, a class-unconditional ImageNet diffusion model at a resolution of 256x256 (Dhariwal & Nichol, 2021) was used as the diffusion model the DDS method.[2]

In the Transformer architecture, interpretability methods obtain the intermediate self-attention matrices of the model for an input and aggregate them alongside heads and layers using various components of the model. The pixel relevance scores are used to create attention heat maps of the input image. All interpretability methods included in Hu et al. (2024) (baselines) are used in the study; these include *Raw Attention*, *GradCAM*, *Rollout*, *LRP* and *Transformer Attribution* (TA). As mentioned, *Attribution Rollout*

[1] Obtained from `https://github.com/huggingface/pytorch-image-models/releases/download/v0.1-vitjx/jx_vit_base_p16_224-80ecf9dd.pth`
[2] Obtained from `https://github.com/openai/guided-diffusion`

was also investigated and implemented according to (Xu et al., 2022). Default hyperparameters were used for all of the methods.

The DDS method was implemented following the algorithm description included in the Appendix of Hu et al. (2024). That is, to convert a ViT into an FViT, Gaussian noise is added to the input sample. That sample is then denoised using an external model (in our case, a guided diffusion model). After this process, the blurred and de-blurred sample is passed to a ViT of choice.

The process of finding the faithfulness region of an FViT extends the typical process found in explainable ViTs by taking the average of multiple samples and performing additional computations. Following the demo implementation of Hu et al. (2024), the number of samples created was 10 for the qualitative images or 2 when an adversarial attack was performed.

4.2 Metrics

In the scope of the reproduction as well as in the extensions, we make use of several performance metrics which are defined as follows.

Pixel Accuracy (Pix. Acc.) Defined as the proportion of all true positive predictions TP over the number of pixels in the image n, i.e. $\frac{TP}{n}$.
Mean Intersection over Union (mIoU) Defined as the mean of the ratios of intersection and union for each of the n_{class} classes, i.e. $\frac{1}{n_{class}} \sum_{i=1}^{n_{class}} \text{IoU}_i$. In our experiments, we use $n_{class} = 2$ with the classes representing the fore- and background.
Mean Average Precision (mAP) Defined as the mean average precision of the segmentation.

4.3 Datasets

The ImageNet-segmentation subset (Guillaumin et al., 2014) [3] was used for the image segmentation task. The subset contains images sourced from the full ImageNet dataset with ground-truth segmentations. Therefore, with a model pre-trained for ImageNet classification, there is no need for fine-tuning or data splitting. The subset consists of 4276 images encompassing 445 classes and 6225 segmentations. The segmentations are binary and only distinguish the foreground from the background. COCO (Lin et al., 2014) and Cityscape (Cordts et al., 2016) datasets due to a lack of experiment scripts, pre-trained model weights, and computational resources needed to train the ViT and diffusion models.

The ImageNet LSVRC 2012 Validation Set (Russakovsky et al., 2015) was used for the classification under the perturbation task. [4] The set contains 50,000 images with corresponding labels equally distributed from 50 classes. Generating attention heat maps for the full set of images was not feasible due to both computational and time constraints. Therefore, sampling with equal probability weights, no replacement and a preset seed was used. A total of 760 images were sampled from the validation dataset. This means that the results are not fully representative but should still indicate the general trends of the algorithms.

4.4 Hyperparameters

No hyperparameter search was performed for the experiments. Pre-trained ViT and diffusion models were used. Unless specified otherwise, the replicated experiments use the default hyperparameters proposed for DDS, which are a noise level of $\frac{8}{255}$ and 45 backward steps for denoising.

The PGD attack was used on the ViT model in the image segmentation tests; the parameters were left to their default implementation values as used by Hu et al. (2024). Specifically, the maximum perturbation was set to $\frac{8}{255}$, the step size to $\frac{2}{255}$ and the number of steps to 10.

[3] Obtained from `https://calvin-vision.net/bigstuff/proj-imagenet/data/gtsegs_ijcv.mat`

[4] Obtained from `https://academictorrents.com/details/5d6d0df7ed81efd49ca99ea4737e0ae5e3a5f2e5` and `https://image-net.org/data/ILSVRC/2012/ILSVRC2012_devkit_t12.tar.gz`

4.5 Experimental setup and code

Following Hu et al. (2024), Chefer et al. (2021) and Xu et al. (2022), in the image segmentation task, the visualization is considered as a soft-segmentation of the image and compared to the ImageNet ground truth segmentation. Performance is measured by pixel accuracy, obtained after thresholding each visualization by the mean value, mean-intersection-over-union (mIoU), and mean-average-precision (mAP), which uses a soft-segmentation to obtain a score that is threshold-invariant. For all experiments, a random seed of 44 was used.

Again following Hu et al. (2024) and Chefer et al. (2021), the image classification under perturbation tests follow a two-stage setting. First, a model with an interpretability method is used to extract the visualizations from the ImageNet dataset. Then, pixels are masked from an input image and the mean top-1 accuracy of the model is measured. In positive perturbation, pixels are masked from the highest relevance to the lowest, while in negative perturbation from lowest to highest. In positive perturbation, a decrease in performance is expected, while in negative perturbation accuracy should be maintained. Both tests used perturbation in the range of 10 to 90% of the pixels in 10% increments. The area-under-the-curve (AUC) of top-1 accuracy over perturbation steps was calculated. DDS was not used in the second stage, i.e. only in the first, due to computational constraints.

Our code is available at github.com/REDACTED.

4.6 Computational requirements

The experiments were run on an NVIDIA A100 GPU partitioned into two instances using Multi-Instance GPU technology, effectively utilising half of the GPU. In addition, 9 cores of Intel Xeon CPUs were utilised.

Table 1 presents the overview of the computational requirements for the segmentation and classification tasks. Baselines + DDS values are missing due to cluster malfunction not reporting the final scores.

Experiment	**Method**	**Wall Time** (hh:mm)	**CPU Util.** (hh:mm)	**CPU Eff.** (%)
Segmentation	Raw Attention	00:59	04:45	53.38
	Raw Attention + DDS	-	-	-
	Rollout	00:57	04:43	55.53
	Rollout + DDS	-	-	-
	GradCAM	00:48	04:35	62.68
	GradCAM + DDS	-	-	-
	LRP	00:57	04:43	55.46
	LRP + DDS	-	-	-
	TA	00:57	04:43	55.53
	TA + DDS	10:08	15:45	17.26
	Attr. Rollout	00:58	04:51	55.40
	AR + DDS	10:06	15:44	17.30
Classification	Raw Attention	00:03	00:22	87.18
	Rollout	00:03	00:22	87.33
	GradCAM	00:01	00:07	68.98
	LRP	00:03	00:23	86.97
	TA	00:03	00:26	86.10
	TA + DDS	08:17	11:45	15.75

Table 1: Computational requirements of the experiments run

5 Results

Briefly, our results for the image segmentation task exhibit the same general trends as in Hu et al. (2024), thereby confirming Claim 1 in Section 2. Namely, explanations obtained from an FViT are indeed more robust under attack compared to regular ViT. However, DDS and some baseline methods perform worse than in the original study. Instead, our results are more consistent with those from Chefer et al. (2021).

Applying DDS to Attribution Rollout substantially improved the performance of Attribution Rollout and also achieved the best overall results in image segmentation. Applying DDS to the baseline interpretability methods saw an improvement only for LRP with no change for the other methods, thereby somewhat confirming the plug-in nature of DDS. Lastly, applying DDS to a ViT resulted in a substantial adverse increase in the environmental impact.

5.1 Results reproducing the original study

Table 2 shows the results of the interpretability methods evaluated in the image segmentation task under a PGD attack.

Method	**Pix. Acc.**		**mIoU**		**mAP**	
	Ours	Hu et al.	Ours	Hu et al.	Ours	Hu et al.
Raw Attention	0.64	0.65	0.43	0.54	0.78	0.82
Rollout	0.72	0.67	0.52	0.56	0.82	0.84
GradCAM	0.67	0.69	0.42	0.58	0.72	0.86
LRP	0.66	0.71	0.40	0.6	0.64	0.88
TA	0.73	0.73	0.52	0.62	0.82	0.9
TA + DDS	**0.77**	**0.76**	**0.58**	**0.65**	**0.85**	**0.93**

Table 2: Reproduced and original pixel accuracy, mIoU, and mAP results from the ImageNet segmentation task under a PGD attack

Same as Hu et al. (2024), the addition of DDS improved TA over all the metrics; however, we found our baselines often did not match the results presented in the original study. For example, *GradCAM* results for mIoU range from 0.42 (ours) to 0.58 (Hu et al., 2024) and for mAP from 0.72 (ours) to 0.86 (Hu et al., 2024). In fact, for each of the methods, at least one of the metrics in our replication is worse compared to the original study.

Furthermore, the ordering in Hu et al. (2024) is unanimously improving: each method improved over worse ones in all metrics in steady increments. However, our results across methods were less distinct: for example, while *TA* had a higher pixel accuracy and mAP than *Rollout*, it had a lower mIoU.

Lastly, same as Hu et al. (2024), our study obtained higher metric values with some methods under attack than Chefer et al. (2021) without attack. A comparison of some metrics obtained for *TA* and *GradCAM* across the two studies and ours can be seen in Table 3.

Method	**Attack**	**Pix. Acc.**	**mIoU**	**mAP**
TA	None (Chefer et al., 2021)	**0.8**	**0.62**	0.86
	PGD (Hu et al., 2024)	0.73	0.62	***0.9***
	PGD (ours)	0.73	0.52	0.82
GradCAM	None (Chefer et al., 2021)	0.64	0.41	0.72
	PGD (Hu et al., 2024)	***0.69***	***0.58***	***0.86***
	PGD (ours)	*0.67*	*0.42*	0.72

Table 3: Results from the ImageNet segmentation task for two methods from Chefer et al. (2021), Hu et al. (2024), and our study. Bold values represent the best metrics for a method across the studies, while italicized represent an increase in a metric after an attack.

Table 4 shows the results of the interpretability methods evaluated in the image classification task under positive and negative perturbation. As mentioned before, the results are on a portion of the original dataset and are therefore not fully representative. Nonetheless, they align with those of Chefer et al. (2021): TA, Rollout and Raw Attention performed well while LRP and GradCAM did not. However, DDS with TA performed comparably well to TA and we were therefore not able to reproduce Hu et al. (2024)'s DDS and TA state-of-the-art results. Furthermore, their ordering and the individual values of some baselines substantially differ from ours: they found that Rollout and Raw Attention performed the worst while LRP and GradCAM did better.

Method	**Pos. Pert.**	**Neg. Pert.**
GradCAM	0.354	0.428
LRP	0.331	0.444
Raw Attention	0.219	0.503
Rollout	0.154	0.601
TA + DDS	0.156	0.602
TA	**0.135**	**0.606**

Table 4: The AUC values of the interpretability methods in the image classification task under positive and negative perturbation. The AUC of the correctly classified images for all perturbation amounts: from 10% to 90% of the pixels removed in 10% increments. For positive perturbation, a lower AUC is better while for negative a higher AUC is better.

Figure 1 was taken from the original study (Hu et al., 2024) and visualizes the attention heat maps under adversarial corruption. The maps were reproduced by rerunning the authors' qualitative demo with the original hyperparameters and can be observed in Figure 2; as can be seen, the images are very similar to nearly identical and further show that DDS improves interpretability robustness.

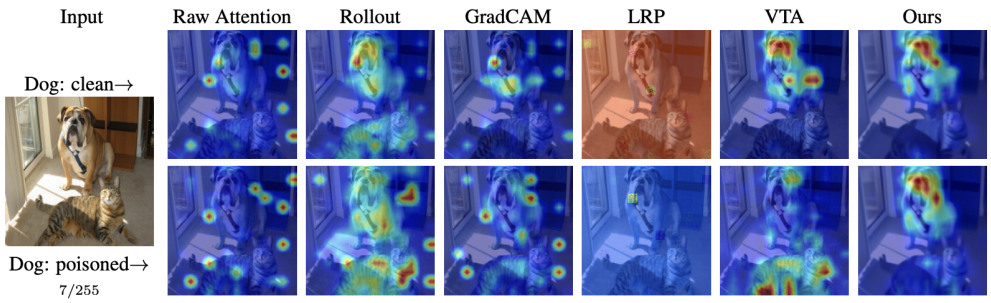

Figure 1: Attention heat maps from the original study.

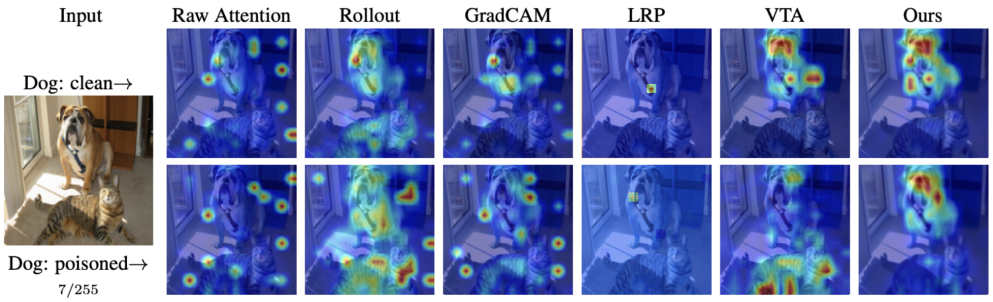

Figure 2: Reproduced attention heat maps.

5.2 Plug-in nature of DDS

The results of applying DDS to the baseline interpretability methods and Attribution Rollout for image segmentation are shown in Table 5. As hypothesized in Section 2: adding DDS improved the performance of Attribution Rollout and also achieved the best overall results in image segmentation. However, with regards to the general plug-in nature of DDS of increasing the robustness of any interpretability method to adversarial attacks, there was no improvement for GradCAM, Rollout and Raw Attention. Adding DDS to LRP, however, yielded improved results, matching the performance of TA. Since LRP, TA and AR all use backwards propagated relevance scores (or attribution scores for Attribution Rollout) as opposed to the remaining methods, DDS may be a suitable method of improving interpretability methods relying on such metrics.

Method	Pix. Acc.	mIoU	mAP
Raw Attention	0.64	0.43	0.78
Raw Attention + DDS	*0.65*	0.43	0.78
Rollout	0.72	0.52	0.82
Rollout + DDS	0.72	0.52	0.82
GradCAM	0.67	0.42	0.72
GradCAM + DDS	0.67	0.42	*0.73*
LRP	0.66	0.40	0.64
LRP + DDS	*0.71*	*0.50*	*0.81*
TA	0.73	0.52	0.82
TA + DDS	*0.77*	*0.58*	*0.85*
AR	0.74	0.53	0.82
AR + DDS	**0.78**	**0.60**	**0.86**

Table 5: Extended pixel accuracy, mIoU, and mAP results from the ImageNet segmentation task under a PGD attack. Italicized values represent improved metrics, while bold ones represent best overall score.

5.3 Environmental impact

One concern that was observed during our preliminary experiments was the computational cost of the FViT achieved through Denoised Diffusion Smoothing. Therefore, we found it appropriate to measure the environmental impact of each interpretability method by reporting its runtime and converting it using the tool presented by Lacoste et al. (2019) with the average carbon efficiency in the Netherlands for 2024 of 370 g eqCO_2/kWh to validate Impact 1[5]. The results are shown in Table 6. In short, applying DDS had a strongly negative impact: applying it to each of the interpretability methods increased the runtime and emissions over 10 times.

6 Discussion

The results confirm that applying DDS produces more robust and stable explanations in the case of a ViT and the ImageNet segmentation task. We have confirmed that the negative effect of a PGD attack on the input sample is partially negated by the technique suggested in Hu et al. (2024). Unfortunately, improvements by applying DDS were not found in the classification under the perturbation task. Furthermore, we do not agree with additional statements from the authors; namely that DDS is efficient and promising for real-world applications with large-scale data. With an increase of the computational time by a factor of 10.7 it is hard to argue for the efficiency of the proposed method, which increases the pixel accuracy by 4% under PGD attack and possibly less under more desirable circumstances in which no adversarial attack takes place.

Our reproduction of the results found in the study by Hu et al. (2024) had several limitations. Namely, our reproductions were limited to ImageNet experiments, due to us not having access to a pre-trained ViT nor the

[5]Obtained with *nowtricity* in January of 2025.

Method	**Execution Time** (s)	**Energy** (J)	**CO_2 Emitted** (g)
Raw Attention	3554	355400	37.53
Raw Attention + DDS	36266	6001885	616.86
Rollout	3390	561333	57.69
Rollout + DDS	36316	6560805	674.30
GradCAM	2912	296822	30.51
GradCAM + DDS	36238	6376733	655.39
LRP	3385	533882	54.87
LRP + DDS	35236	6220494	639.33
TA	3388	956766	98.33
TA + DDS	36494	4577928	470.51
AR	3488	430928	44.29
AR + DDS	36341	5316731	546.44

Table 6: Environmental impact of one validation run of each ViT explainability method

exact methodology of approaching segmentation and perturbation tasks on COCO and CityScape datasets. Additionally, we were forced by technical limitations to investigate the classification under perturbation task only on a subset of ImageNet. We believe both of those limitations could be addressed by further research.

We will now go into more detail regarding reproducing the results from Hu et al. (2024)'s article.

6.1 Reproducibility challenges

We found it to be unproblematic to access the baseline methods' implementations discussed in Chefer et al. (2021) as well as Attribution Rollout (Xu et al., 2022) and incorporate it alongside DDS for the ImageNet segmentation test. Additionally, installation of the working environment (on a Linux-based workspace) proved to be fairly straightforward, making the reproduction of the qualitative demo very easy.

However, we encountered multiple hardships in reproducing the results from Hu et al. (2024). Firstly, an official repository with the code was not listed in the article at the time of writing this and we had to base our implementations on a repository by one of the authors containing only a demo Jupyter notebook. We later learned that this repository contained outdated implementations.

Secondly, we found multiple discrepancies between the demo notebook and the information specified in the article. One example of this is a demo notebook varying the noise amount between cells, while the article specified a specific number.

As we lacked a full implementation of the experimental setup, we also found the information provided in the article to be incomplete and we had to resort to assumptions made based on the limited implementations in the demo. This includes reporting classification accuracy as part of results in Table 1 (the equivalent of Figure 2 in this paper), which is never defined in the methodology, hence why it is omitted in our work. In addition, the number of input samples to be processed with DDS and subsequently combined was not precisely defined, leading us to manually specify the value of 2 for the two tasks. This has an impact on the computational cost and energy consumed. Nonetheless, the conclusions regarding the environmental impact should still stand with a lower value (1).

Furthermore, some methods (taken from Chefer et al. (2021) and later used by Hu et al. (2024) and Xu et al. (2022)) were not set up in a computationally efficient manner: for example, the diffusion denoising process was set up to support a batch size of only 1, which resulted in the potential underutilization of GPUs and therefore longer runtimes.

Lastly, experiment specifications of the COCO and Cityscape datasets, the details of the segmentation and classification tasks and how the corresponding ViTs are trained are lacking. This led us to focus on the ImageNet tasks exclusively.

6.2 Communication with original authors

We contacted the authors by email, mostly regarding the implementation details. The authors made us aware that the code base that was published, which we have used to reproduce the results, is an outdated version and that the paper version will be published soon. However, the paper version was never made available to us.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020. URL https://arxiv.org/abs/2005.00928.

Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016. URL https://arxiv.org/abs/1604.00825.

Houssem Ben Braiek and Foutse Khomh. Machine learning robustness: A primer, 2024. URL https://arxiv.org/abs/2404.00897.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, June 2021.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2021. URL https://arxiv.org/abs/2012.00364.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110:328–348, 2014.

Lijie Hu, Yixin Liu, Ninghao Liu, Mengdi Huai, Lichao Sun, and Di Wang. Improving interpretation faithfulness for vision transformers. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=YdwwWRX20q.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, 2020. URL https://arxiv.org/abs/2004.03685.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Li Xu, Xin Yan, Weiyue Ding, and Zechao Liu. Attribution rollout: a new way to interpret visual transformer. *Journal of Ambient Intelligence and Humanized Computing*, 14:163–173, 2022. URL https://api.semanticscholar.org/CorpusID:251942826.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2021. URL https://arxiv.org/abs/2012.15840.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. URL https://arxiv.org/abs/2010.04159.