RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] On the Reproducibility of CartoonX

Elias Dubbeldam[1], Aniek Eijpe[1], Jona Ruthardt[1], and Robin Sasse[1]
[1]University of Amsterdam, Amsterdam, The Netherlands – [1]Equal contribution

Edited by
Koustuv Sinha,
Maurits Bleeker,
Samarth Bhargav

Received
04 February 2023

Published
20 July 2023

DOI
10.5281/zenodo.8173672

Reproducibility Summary

Scope of Reproducibility – CartoonX [1] is a novel explanation method for image classifiers. In this reproducibility study, we examine the claims of the original authors of CartoonX that it: (i) extracts relevant piece-wise smooth parts of the image, resulting in explanations which are more straightforward to interpret for humans; (ii) achieves lower distortion in the model output, using fewer coefficients than other state-of-the-art methods; (iii) is model-agnostic. Finally, we examine how to reduce the runtime.

Methodology – The original authors' open-sourced implementation has been used to examine (i). We implemented the code to examine (ii), as there was no public code available for this. We tested claim (iii) by performing the same experiments with a Vision Transformer instead of a CNN. To reduce the runtime, we extended the existing implementation with multiple enhanced initialization techniques. All experiments took approximately 38.4 hours on a single NVIDIA Titan RTX.

Results – Our results support the claims made by the original authors. (i) We observe that CartoonX produces piece-wise smooth explanations. Most of the explanations give valuable insights. (ii) Most experiments, that show how CartoonX achieves lower distortion outputs compared to other methods, have been reproduced. In the cases where exact reproducibility has not been achieved, claim (ii) of the author still holds. (iii) The model-agnosticism claim still holds as the overall quality of the ViT-based explanations almost matches that of the CNN-based explanations. Finally, simple heuristical initializations did not improve the runtime.

What was easy – The mathematical background and intuition of CartoonX were clearly explained by the original authors. Moreover, the original author's code was well structured and documented, which made it straightforward to run and extend.

What was difficult – Some hyperparameter settings and implementation details needed to reproduce the experiments were not clear or transparent from the original paper or code. This made it difficult to implement and reproduce these experiments.

Communication with original authors – We have been in brief communication with the original authors. They were able to address most of our points, providing us with some additional clarifications about the exact implementation and hyperparameter settings.

Correspondence should be addressed to Elias Dubbeldam (elias.dubbeldam@student.uva.nl)
The authors have declared that no competing interests exist.
Code is available at https://github.com/JonaRuthardt/MLRC-CartoonX. – SWH swh:1:dir:691b7400b38017cfe651a59735dc715755388c4d;origin=https://github.com/JonaRuthardt/M
CartoonX;visit=swh:1:snp:ffead088e1c166ac58b0f37c5c8e87344f638b3d;anchor=swh:1:rev:38364e7eef0fbc99f2b143fad90a29dae113f95c.
Open peer review is available at https://openreview.net/forum?id=MK4IQJdLLeo.

1 Introduction

The trend towards using ever-more complex machine learning models in the field of Computer Vision has led to improved accuracy at the cost of interpretability. Most state-of-the-art models are inherently opaque, making understanding their inner dynamics and decision-making processes difficult. This has motivated the emerging research area of explainable AI [2], with a strong focus on explaining the classification of images. Numerous explanation methods, like Smoothgrad [3] or LIME [4], have been developed for image classifiers. These methods all operate in the pixel domain, producing pixel-sparse or jittery explanations.

Challenging this approach, the authors of [1] proposed CartoonX; a novel, model-agnostic explanation method for image classifiers that operates in the wavelet domain. This paradigm shift is motivated by the idea that demanding sparsity in the wavelet domain introduces piece-wise smooth explanations, i.e., asking the question *What is the piece-wise smooth part of the input signal that leads to the model decision?* [1]. CartoonX generates explanations by applying the rate-distortion explanation (RDE) framework (originally proposed by [5] and extended by [6]) in the wavelet domain of an image. RDE is an optimization problem that enforces maximum sparsity with minimum distortion in the model's output. This is achieved by optimizing a deletion mask applied to the image coefficients, thus marking relevant components. In the conventional setting, Pixel RDE enforces sparsity on (super-)pixels. In the case of CartoonX, RDE is used to enforce sparsity across the wavelet coefficients.

In this contribution, we aim to reproduce the results of the CartoonX paper and perform an additional experiment to examine the authors' main claims. Moreover, an extension is proposed to improve the runtime of CartoonX.

The remainder of this paper is organized as follows: Section 2 presents the scope of this reproducibility study and the methodology is outlined in Section 3. In Section 4, the results are presented and discussed. Section 5 discusses different aspects of the reproducibility effort. Lastly, in Section 6, a conclusion is given.

2 Scope of Reproducibility

In this reproducibility study, we examine the original authors' claims:

- CartoonX extracts relevant piece-wise smooth parts of the image, resulting in more straightforward explanations.
- CartoonX achieves lower distortion in the model output while using fewer coefficients than other state-of-the-art methods.
- CartoonX is model-agnostic.

The explanations provided by CartoonX are qualitatively evaluated (i.e., manually compared and interpreted) to examine the first claim. To investigate whether CartoonX achieves lower distortion in the model output, the distortion of the different methods is compared. This is referred to as the quantitative evaluation.

The original authors used two CNNs in their experiments to investigate their claim of model agnosticism. We extend their experiments by running CartoonX with a Vision Transformer (ViT). The results of CartoonX with a ViT are compared to an attention map of the Transformer model. Attention maps have been used in other works to explain model decisions [7], serving as a basis for the cross-validation of CartoonX.

The original authors suggested to train a neural network to predict a good initialization of the deletion mask for arbitrary images of the target distribution with the intention of significantly reducing CartoonX's runtime. This should ideally lead to faster convergence. Due to the computational cost of generating sufficient training data, two different heuristical strategies were assessed.

Summarizing, the main contributions of this publication are:

- Examining the first two claims of the original authors by reproducing the experiments and qualitative and quantitative evaluating the results.
- Applying CartoonX with a ViT to investigate the model agnosticism claim.
- Exploring using simple heuristics for initialization to improve runtime.

3 Methodology

The original authors provide a publicly available, well-documented, and cleanly written codebase,[1] containing all necessary implementations to produce CartoonX and Pixel RDE explanations. Nonetheless, neither an implementation for their quantitative evaluations nor a reference to the original images was published, complicating the reproducibility effort. The authors' implementation was used as a baseline and adapted to accommodate the outlined extensions and it was supplemented with the quantitative experiments. We consulted with the authors to verify the correctness of our interpretation. The extended open-source repository is made publicly available.[2]

3.1 CartoonX

CartoonX [1] is a rate-distortion-based explanation method for image classifiers operating in the wavelet domain to identify the components (i.e., wavelet coefficients) of an image that are most decisive for the model's prediction.
For this, a given image is transformed into its wavelet representation by applying the Discrete Wavelet Transform (DWT). Out of the resulting DWT coefficients h that represent the image in wavelet space, the least relevant components are masked. This is done by iteratively learning a mask s that minimizes a distortion metric while encompassing minimal components. The sparsity is enforced by applying the ℓ_1-norm on the mask's values and controlled using an additional parameter[3] λk by which this loss component is multiplied. Starting with an all-ones initialization, the mask's values are continuously decreased for wavelet components that contain little classification decisive information. At every iteration, a batch of L adaptive Gaussian noise samples $v^{(1)}$, ..., $v^{(L)}$ is drawn. With these samples, the obfuscations $y^{(1)}$, ..., $y^{(L)}$ are computed as $y^{(i)} = \text{DWT}^{-1}(h \odot s + (1-s) \odot v^{(i)})$. Therefore, less relevant components with mask values close to 0 will be (partially) replaced by noise. The efficacy of the mask is ascertained by computing the distortion $\hat{D}(x, s, \Phi)$ as the average squared distances of the post-softmax originally predicted class probabilities of the original image x and the set of obfuscations. Together with the sparsity constraint, the loss objective emerges as $l(s) = \hat{D}(x, s, \Phi) + \lambda\|s\|_1$.
The explanation is ultimately obtained by applying the learned mask to the image's wavelet coefficients and converting the resulting representation back into pixel space as a grayscale image. This results in a piece-wise smooth image explaining the model's decision by highlighting relevant areas for each image. For a more comprehensive and conclusive background on the rate-distortion framework and the exact implementation of CartoonX, we refer back to Sections 3 to 5 in [1].

3.2 Model descriptions

To reproduce the results from the original paper, we also used a pre-trained MobileNetV3-Small [8] (Top-1 accuracy 67.7%; 1.8M parameters). It was additionally

[1] https://github.com/skmda37/CartoonX

[2] https://github.com/JonaRuthardt/MLRC-CartoonX

[3] Parameter k refers to the number of pixels, while λ refers to the sparsity level. We will treat them as a single parameter to retain consistency with the original authors.

used to test for the speed-up effects of different mask initialization techniques. Furthermore, a pre-trained Transformer-based classifier in the form of DeiT-tiny [9] (Top-1 accuracy 72.2%) was examined. This particular ViT was chosen due to its comparatively few 5M parameters, thus, behaving runtime efficiently. All models used in this study and the original paper were pre-trained on ImageNet1K.

3.3 Datasets

For all experiments, we used the same random sub-set of 100 images of 100 distinct but randomly selected classes from ImageNet.[4] In line with the original publication, the images were resized to 256×256 pixels. Only for experiments involving the Transformer-based model (i.e., the *Model Agnosticism Experiment*) were the images resized to 224×224 pixels to ensure model compatibility.

3.4 Experimental setup and code

In order to verify and extend the claims made in [1], three different experimental setups are proposed and specified in this Section.

Reproducibility Experiment — The reproduction of the experiments consists of two parts. The qualitative experiment, corresponding to the claim that CartoonX is qualitatively better to interpret, and the quantitative experiment, corresponding to the claim that CartoonX achieves lower distortion while using fewer coefficients. Both quantitative and qualitative experiments were evaluated akin to the original paper.
For the qualitative experiment, explanations for the 100 images with both the CartoonX and Pixel RDE methods were created. The most insightful and interesting explanations are used to highlight and discuss the interpretability of CartoonX compared to Pixel RDE. To increase transparency and mitigate potential selection biases, all results underlying the qualitative evaluations are made publicly available in our repository.
The quantitative experiment consists of three different subexperiments. In the first two subexperiments, the optimized masks for the 100 explanations of the qualitative experiment are used. For the first subexperiment, all components are randomized with adaptive Gaussian noise, except for an iteratively increasing fraction of the most relevant components, (i.e., the highest mask values). Conversely, in the second experiment, the most relevant components are randomized. Finally, the third subexperiment examines the distortion and non-sparsity (the two loss terms) for varying λk.

Model Agnosticism Experiment — To examine the claim that CartoonX is model-agnostic, the ViT DeiT-tiny [10] was integrated into the CartoonX framework. For all 100 images, three different explanations were created: a CartoonX explanation for both the ViT as well as the CNN, and the attention rollout [11]. The latter method linearly combines the attention weights throughout the layers of the vision transformer. More specifically, at each layer, it merges the attention at each position with the attention at each position of the previous layers. To account for the multiple attention heads, we take an average of all heads. The implementation by [12] is used to create the attention rollout. Moreover, the quantitative evaluation for this experiment is set up analogously to the quantitative evaluation for the reproducibility experiment.

Runtime Efficiency Experiment — To improve runtime, we explore using simple heuristics for initializing the deletion mask. Two different strategies were tested. In the first strategy, we use an efficient preoptimization algorithm, which iteratively decreases the initial mask from 1 to 0 in one-percent increments until the network predicts a new class.

[4] https://github.com/EliSchwartz/imagenet-sample-images

This pre-convergence criterion was chosen heuristically. Up until this point, the gradients are still largely useful. The second strategy uses a binary foreground mask as an initialization, with 1 for wavelet coefficients corresponding to foreground regions and 0 for those corresponding to background regions. The main idea is that the network primarily uses the foreground to predict the class. Consequently, it is the part of the image containing most of the relevant frequencies needed to explain the model's decision. The efficacy of each of these approaches was evaluated based on the loss curves obtained during the actual optimization. Hence, it is possible to ascertain that the model converges towards approximately the same loss value and if it does so after fewer iterations. We compared these strategies to a random initialization and the default all-ones initialization, as implemented by the original authors.

3.5 Hyperparameters

For all experiments, the default values for all hyperparameters – as specified in [1], Section 5 – were used. Whenever the hyperparameters were not specified, we used the default values in the code provided by the original authors. To choose a proper value for λk for the ViT, we did a qualitative search (see Appendix A for details). For the attention rollout explanation, the attention heads were fused by taking the mean (as done in [11]) and the discard ratio of 0.9 is chosen, to focus on the highest attention values. The following table gives an overview of the experimental setups used:

Experiment	CNN λk	ViT λk	P. RDE λk	iter.	b. size	optimizer	lr	init. mask
Reprod.	20	N/A	4	2000	64	Adam [13]	10^{-3}	ones
Agnost.	20	10	N/A	2000	64	Adam	10^{-3}	ones
Runtime	20	N/A	N/A	2000	64	Adam	10^{-3}	various

Table 1. Overview of hyperparameter settings used in our experiments.

3.6 Computational requirements

Obtaining a singular explanation with the Mobilenet-based CartoonX and Pixel RDE approaches required 96 and 75 seconds on an NVIDIA Titan RTX, respectively. This equates to a total GPU walltime of 28.5 hours to obtain all quantitative and qualitative results of 100 images with six different λk. This is proportional to the reported times in [1]. Furthermore, the creation of all relevant explanations for the model agnosticism experiment requires 272 seconds per image. Thus, for this whole experiment, 7.5 hours of runtime were used. Finally, to test various mask initializations, 2.3 hours were used. The overall GPU utilization during this study amounted to approximately 38.4 hours.

4 Results and Discussion

We performed a qualitative and quantitative analysis of the results of CartoonX and compared them to Pixel RDE. Furthermore, CartoonX's performance on ViTs was evaluated and the results were compared to the corresponding attention masks. Lastly, different initialization strategies for the deletion mask, including a preoptimization algorithm intended to improve the runtime of CartoonX, were examined.

4.1 Results reproducing original paper

Qualitative Reproducibility Experiment — The original authors claimed that CartoonX extracts relevant piece-wise smooth parts of the image, resulting in more intuitive explanations. In Fig. 1 we present a selection of explanation comparisons between CartoonX and Pixel RDE. Pixel RDE produces pixel-sparse explanations. Conversely, CartoonX introduces sparsity in the wavelet domain, blurring out irrelevant areas of the image. These characteristics are what the notion of piece-wise smooth areas refers to.

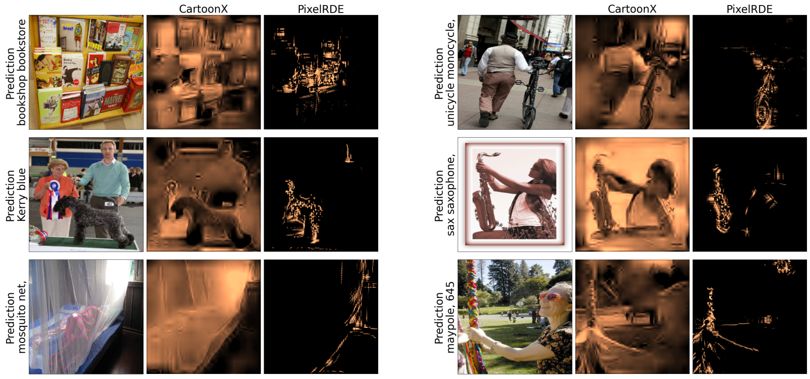

Figure 1. Examples of images where CartoonX produces subjectively better explanations (left) and examples of images where Pixel RDE produces subjectively better explanations (right).

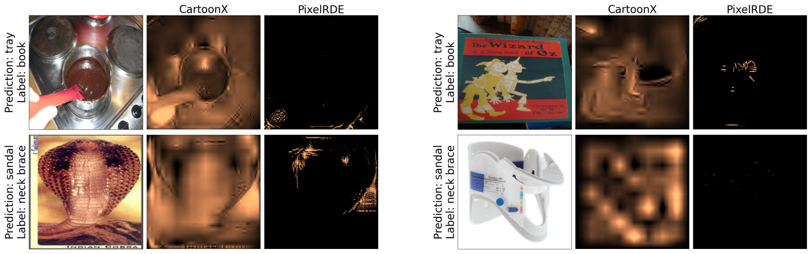
CartoonX
PixelRDE
Prediction: tray Label: book
Prediction: sandal Label: neck brace
CartoonX
PixelRDE
Prediction: tray Label: book
Prediction: sandal Label: neck brace

Figure 2. Examples of misclassifications where CartoonX provides a useful explanation for further investigation (left) and where CartoonX fails at providing a useful explanation (right).

Fig. 1 highlights selected samples where either CartoonX outperforms Pixel RDE (left) or vice-versa (right). Here, *outperform* refers to the subjective, qualitative evaluation of the results. In the left images, the explanations provided by Pixel RDE are sparse and hard to interpret. Sometimes irrelevant parts of the background have also been marked. CartoonX conserves the shape and, to some degree, the texture of the relevant parts of the image. In the right images, all objects themselves are identified as the principal explanations by Pixel RDE. CartoonX also considers the people using these things and part of the background as an explanation. Nonetheless, overall it was much easier to find examples of CartoonX outperforming Pixel RDE than the other way around. Even in the latter case, CartoonX still provides predominantly reasonable explanations.
Fig. 2 shows instances where the model fails and CartoonX either indicates potential reasons (left) or fails to deliver an interpretable explanation (right). The left images show how the outlines of the objects resemble objects of other classes, giving engineers a chance to adapt their models. Pixel RDE cannot produce this explanation. In the right images, neither method provides a useful explanation.

Quantitative Reproducibility Experiment — The original authors claimed that CartoonX achieves lower distortion in the model output while using fewer coefficients than other state-of-the-art methods. Fig. 3 shows three qualitative evaluations of CartoonX vs. Pixel RDE. The left-most plot depicts the rate-distortion curve when keeping the most relevant coefficients while randomizing the others. The relevance corresponds to the associated mask value. A good explanation yields a steep decrease in the distortion for low rates, as few coefficients are necessary to classify the image consistently. The middle plot shows the rate-distortion curve when randomizing the most relevant coefficients while keeping the others. A good explanation induces a sharp initial increase. The right-most plot shows the distortion as a function of the sparsity-enforcing hyperparameter λk. A suit-

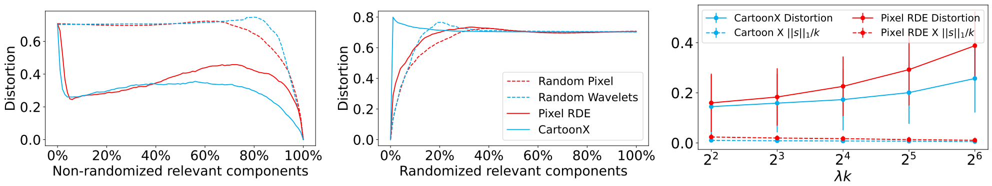

Figure 3. Distortion as a function of relevant components (left, middle), identified by the explainability method. Distortion as a function of the sparsity settings (right).

able explanation constitutes a compromise between low distortion and high sparsity. Across all subexperiments, CartoonX matches or outperforms Pixel RDE.
The exact results diverge in three slight ways from Fig. 7 in the original paper. First, the steep decrease for low rates of non-randomized components for CartoonX and Pixel RDE (solid lines, left plot) differs from the original figure in terms of magnitude (it drops to 0.25 in ours vs. 0.45 in [1]). Even though the drop is steeper in our reproduction, the general result stays the same. They indicate that the most relevant components are equally important for CartoonX and Pixel RDE initially. Second, we observe a slight increase of distortion for both methods after the initial drop, whereas in the original paper the distortion dropped more continuously. Since we mostly care about the relative initial drop, we still confirm the conclusion that CartoonX is superior according to this metric. Third, the obtained non-sparsity values when varying λk (dashed lines, right plot) are significantly lower, despite the curves' general shape being similar. Notwithstanding, our results are in line with the claims made by the original authors, as the non-sparsity value values for Pixel RDE (red) are always higher than for CartoonX (blue).

4.2 Results beyond original paper

Model Agnosticism Experiment — We examine the claim that CartoonX is model-agnostic by running CartoonX on a ViT. Fig. 4 shows four resulting sample explanations, qualitatively comparing the original image, CartoonX for CNN, CartoonX for ViT, and the attention rollout. For a fair comparison, only the cases where the CNN and the ViT predicted the same and the correct class were considered. On the left side in Fig. 4, two examples are shown where both CartoonX for CNN and CartoonX for ViT provide a helpful explanation. Moreover, for both images, this coincides approximately with the attention rollout. On the right side in Fig. 4, two examples are shown where CartoonX for ViT provides a very sparse, almost completely black, explanation while CartoonX for CNN does not. For both images, the attention rollout is not sparse and mainly marks the upper background of the image as having high attention values, also not providing an interpretable reason for the models' decision.
In Fig. 5, both curves for CartoonX (with CNN and ViT) follow a similar shape. The distortion achieved with the ViT drops sharply when randomizing all but the most relevant components (left) and increases sharply when randomizing the most relevant components (right). This is in accordance with the findings for CartoonX on CNNs.
Overall, when regarding the results of all 100 images, the majority of ViT CartoonX explanations are sparser and more sensitive to the choice of λk compared to their CNN counterparts. Furthermore, in Fig. 5, both a sharper initial drop (left) and increase (right) can be observed for the CNN compared to the ViT, indicating marginally worse performance for ViTs. While the CNN-based explanations are marginally superior, the correct identification of relevant components in the ViT case is still apparent. Hence, our findings mostly support the claim of model agnosticism.

Runtime Efficiency Experiment — The original authors proposed using neural networks to predict an initialization for the deletion mask to speed up their algorithm's runtime. We

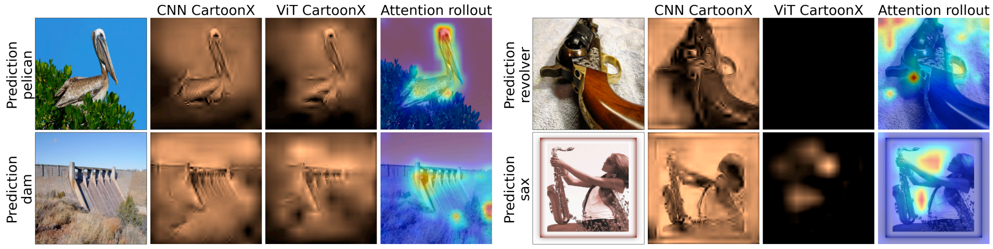

Figure 4. Cases where CartoonX provides a useful explanation for both CNN and ViT (left) and cases where CartoonX (ViT) and the attention rollout do not provide an intelligible explanation of the model's decision (right).

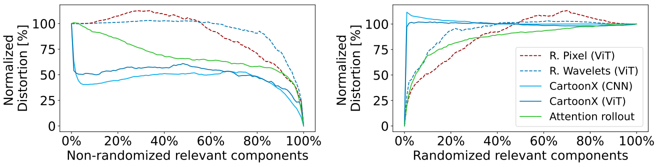

Figure 5. Normalized distortion as a function of relevant components, identified by CartoonX and the attention rollout (random pixels/wavelet coefficients, and attention rollout as benchmark).

investigated whether simple heuristics could already produce a good initialization suitable for that purpose.

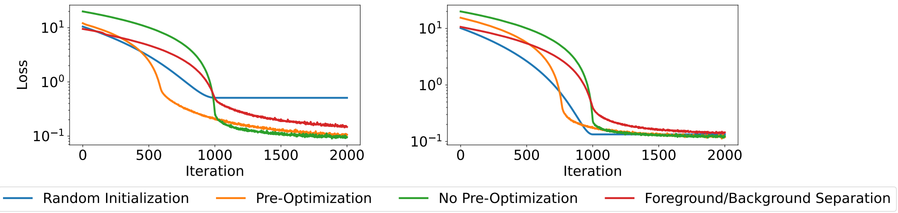

Figure 6. Comparing the loss curves of the CartoonX optimization algorithm with different deletion mask initializations for two different images.

Fig. 6 shows the loss curves for different initialization strategies for two sample images. While the preoptimized mask leads to faster initial convergence, the loss curve flattens quickly. Before qualitatively good convergence, the preoptimization and normal initialization curves reunite again. It is important to note that the slight differences in loss after around 1000 iterations make a notable difference in the explanation's quality. The foreground segmentation did not yield beneficial results due to the final loss being too high compared to other methods. Lastly, random initialization was tested, which led to unrobust results.

These outcomes indicate that more complex approaches might be required to obtain the desired speedup. Such an approach could be to utilize neural networks to predict the initial deletion mask. Furthermore, our results suggest that these networks must act risk averse, i.e., using a less sparse mask rather than blocking out many wavelet coefficients. The reason for that is that unrobust results were observed for any mask, which was already made too sparse at initialization.

4.3 Critique of our methods

For all experiments, a value for λk was qualitatively, thus somewhat subjectively, chosen. Being limited to running the experiments on a small subset of ImageNet, consisting of 100 random images from distinct classes, the samples were not entirely randomly chosen (no duplicate classes were included). Nonetheless, this method ensures more diversity, especially in the qualitative analysis. The lack of a decisive measure to define convergence complicated the determination of a suitable number of training iterations. Furthermore, it led to a rather vague interpretation of what can be considered a speedup of the algorithm. Lastly, it should be noted that the attention mask, used for comparison in the ViT experiment, is not explicitly designed to serve as an explanation [14].

5 Reproducibility review

5.1 What was easy

The original paper had an extensive explanation of the background of CartoonX, both mathematically and intuitively. Moreover, they provided an article on wavelets[5] for explainability, making it easier to understand. The provided implementation was well-documented and ran trouble-free. Therefore, the qualitative experiments were easy to replicate, by merely executing the code on different images and analyzing the results. Furthermore, it was straightforward to extend their code, as it was well-modularized.

5.2 What was difficult

Recreating Fig. 3 was difficult due to uncertainties of which hyperparameters (λk, number of iterations) or models were used. Furthermore, since there is no convergence criterion provided for CartoonX, it was difficult to get an intuition for the loss curves. Lastly, the original paper did not specify the exact subset of ImageNet images. This necessitates a more general evaluation but hinders direct comparisons between the original paper and ours.

5.3 Communication with original authors

We inquired about clarifications on the values used for λk for the qualitative analysis. The authors reported that they used variant values for different images. Nonetheless, only the same value for each image was used in this study to ensure consistency between different images. We further enquired about the λk values used to recreate Fig. 7(a) and (b) of the original paper. Unfortunately, confirmation regarding these values was not provided. Overall, the authors were quick to respond and were open to answer most of the questions as detailed as possible.

6 Conclusion

CartoonX is a valuable explanation method that yields piece-wise smooth explanations. We found this explanation style to be more interpretable than pixel-sparse explanations. It works well for CNNs and, for the most part, also yields good explanations for ViTs. Overall, it is a valuable addition to the ever-growing set of explanation methods available to deep learning researchers, engineers, and users.
Future research could explore how to define a decisive measure of convergence for CartoonX. Such a measure would help evaluate the effectiveness of smart initialization strategies to improve runtime. Specifically, we see potential to investigate neural networks for predicting initial deletion masks, as previously discussed and already suggested by the original authors. Lastly, considering a wider range of image-specific λk values, especially for the ViT, might improve the overall quality of the explanations.

[5] https://julheg.github.io/waveletexplainability/

References

1. S. Kolek, D. A. Nguyen, R. Levie, J. Bruna, and G. Kutyniok. “Cartoon explanations of image classifiers.” In: **European Conference on Computer Vision**. Springer. 2022, pp. 443–458.
2. A. Das and P. Rad. “Opportunities and challenges in explainable artificial intelligence (xai): A survey.” In: **arXiv preprint arXiv:2006.11371** (2020).
3. D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. “Smoothgrad: removing noise by adding noise.” In: **arXiv preprint arXiv:1706.03825** (2017).
4. M. T. Ribeiro, S. Singh, and C. Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier.” In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. 2016, pp. 1135–1144.
5. J. MacDonald, S. Wäldchen, S. Hauch, and G. Kutyniok. “A rate-distortion framework for explaining neural network decisions.” In: **arXiv preprint arXiv:1905.11092** (2019).
6. C. Heiß, R. Levie, C. Resnick, G. Kutyniok, and J. Bruna. “In-distribution interpretability for challenging modalities.” In: **arXiv preprint arXiv:2007.00758** (2020).
7. H. Chefer, S. Gur, and L. Wolf. “Transformer interpretability beyond attention visualization.” In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2021, pp. 782–791.
8. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. “Searching for mobilenetv3.” In: **Proceedings of the IEEE/CVF international conference on computer vision**. 2019, pp. 1314–1324.
9. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. “Training data-efficient image transformers & distillation through attention.” In: **International Conference on Machine Learning**. PMLR. 2021, pp. 10347–10357.
10. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. “Training data-efficient image transformers amp; distillation through attention.” In: **International Conference on Machine Learning**. Vol. 139. July 2021, pp. 10347–10357.
11. S. Abnar and W. Zuidema. “Quantifying attention flow in transformers.” In: **arXiv preprint arXiv:2005.00928** (2020).
12. J. Gildenblat. **Explainability for Vision Transformers (in PyTorch)**. https://github.com/jacobgil/vit-explain. 2020.
13. D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization.” In: **arXiv preprint arXiv:1412.6980** (2014).
14. S. Jain and B. C. Wallace. “Attention is not explanation.” In: **arXiv preprint arXiv:1902.10186** (2019).

A Hyperparameter search

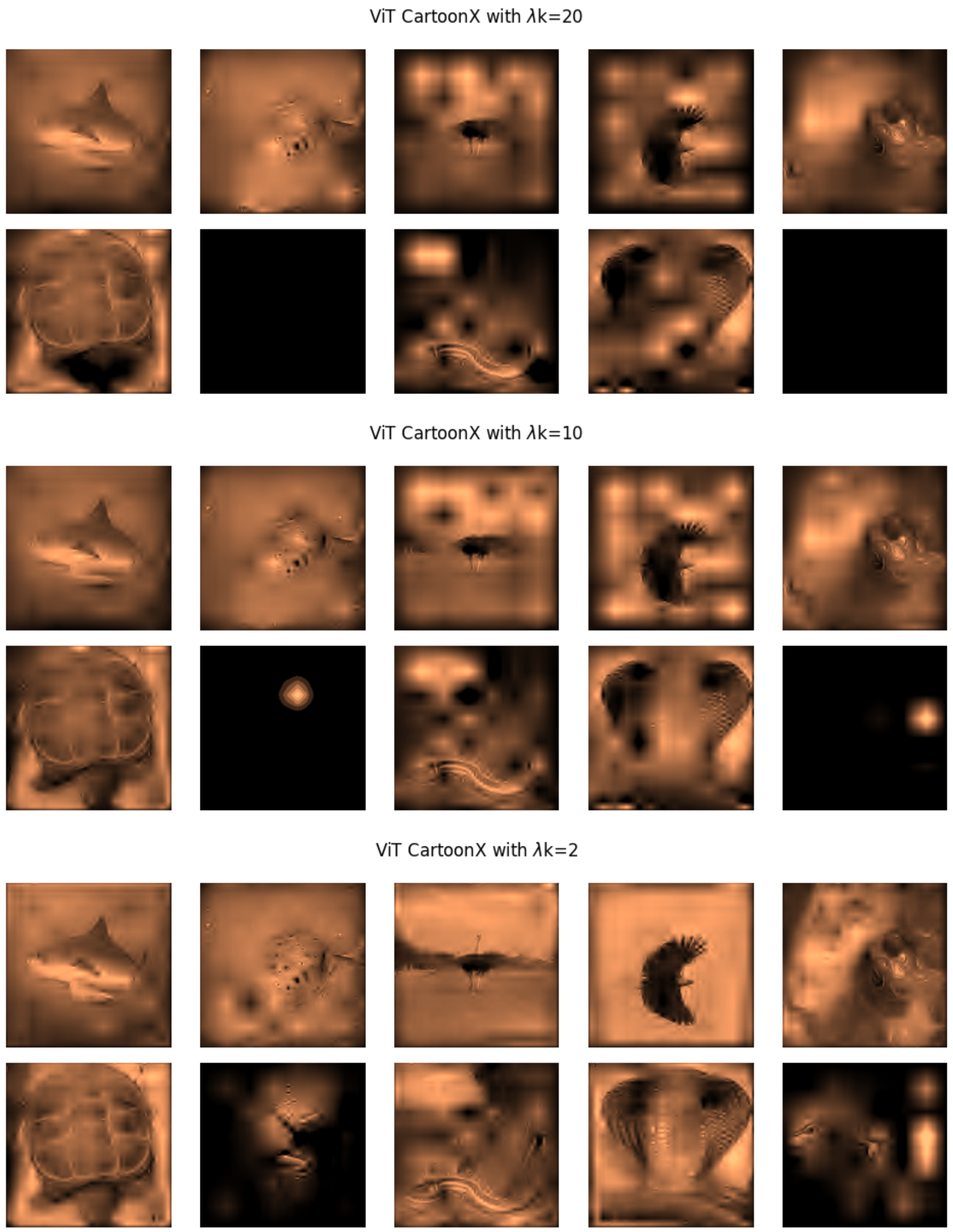

Figure 7. Ten examples of CartoonX with a ViT with $\lambda k = 2, 10, 200$ are depicted in the two top, middle and bottom rows, respectively. Overall, with $\lambda k = 20$, most of the explanations are relatively sparse, with some explanations being completely black. With $\lambda k = 2$ there are no entirely black explanations. However, with this setting some explanations of images did not contain a lot of sparsity, i.e. did not show a clear explanation. Utilising $\lambda k = 10$ constituted a suitable trade-off between the explanations' sparsity and their expressiveness for most images.