Cartoon Explanations of Image Classifiers

Stefan Kolek[1], Duc Anh Nguyen[1], Ron Levie[3], Joan Bruna[4], and Gitta Kutyniok[1,2]

[1] Ludwig Maximilian University of Munich, Germany
{kolek,danguyen,kutyniok}@math.lmu.de
[2] University of Tromsø, Norway
[3] Technion-Israel Institute of Technology, Israel
levieron@technion.ac.il
[4] New York University, NY, USA
bruna@cims.nyu.edu

Abstract. We present *CartoonX* (Cartoon Explanation), a novel model-agnostic explanation method tailored towards image classifiers and based on the rate-distortion explanation (RDE) framework. Natural images are roughly piece-wise smooth signals—also called cartoon-like images—and tend to be sparse in the wavelet domain. CartoonX is the first explanation method to exploit this by requiring its explanations to be sparse in the wavelet domain, thus extracting the *relevant piece-wise smooth* part of an image instead of relevant pixel-sparse regions. We demonstrate that CartoonX can reveal novel valuable explanatory information, particularly for misclassifications. Moreover, we show that CartoonX achieves a lower distortion with fewer coefficients than state-of-the-art methods.

1 Introduction

Powerful machine learning models such as deep neural networks are inherently opaque, which has motivated numerous explanation methods over the last decade (see for example the survey by [4]). A significant fraction of the research literature has focused on explaining image classifications due to both the practical relevance of computer vision tasks and the ease at which heatmaps can communicate explanatory information. Despite the great variety in methods and explanation philosophies, all current methods share the following characteristic: they operate in pixel space. Roughly speaking, existing explanation methods for image classifiers either allocate additive attribution scores to each (super)pixel or optimize a deletion mask on the pixel coefficients to mark a relevant set of pixels. The result is typically a pixel-sparse and jittery explanation. We challenge the conventional approach to explain in pixel space by successfully applying the rate-distortion explanation (RDE) framework [18, 10] in the wavelet domain of images. Our novel explanation method, *CartoonX*, extracts the relevant piece-wise smooth part of an image. Instead of demanding sparsity in pixel space, as in [18, 2], CartoonX demands sparsity in the wavelet domain, which produces piece-wise smooth explanations. Piece-wise smooth images are also known as

cartoon-like images [14]—a class of 2D signals that has been well studied, and for which wavelets provides an efficient representation system [25]. Our work makes the following contributions.

Reformulation and reinterpretation of the RDE framework: We reformulate the RDE framework in a more general manner with enhanced flexibility in the input representation to accommodate complex interpretation queries such as "What is the piece-wise smooth part of the input signal that leads to its model decision?". Thereby, we reinterpret RDE as a simplification of the input signal, which is interpretable to humans and adheres to a meaningful interpretation query. The simplification is achieved by demanding sparsity in a suitable representation system, which sparsely represents the class of explanations that are desirable for the interpretation query.

CartoonX, a novel explanation method tailored to image classifiers: CartoonX is the first explanation method to extract the relevant piece-wise smooth part of an image instead of relevant pixel sparse regions. This is achieved by demanding sparsity in the wavelet domain of images, where sparsity translates into piece-wise smooth images. We demonstrate that our piece-wise smooth explanations can reveal relevant piece-wise smooth patterns that are not easily visible with existing pixel-based methods. Quantitatively, we also corroborate that CartoonX achieves a lower distortion in the model output using fewer coefficients than other state-of-the-art methods.

2 Related Work

The Rate-Distortion Explanation (RDE) framework was first introduced in [18], and extended in [10], as a mathematically well-founded and intuitive explanation framework. RDEs are model-agnostic explanations and inspired by rate-distortion theory, which studies lossy-data compression. An explanation in RDE consists of a relatively sparse mask over the input features, highlighting the relevant set of features. The mask is optimized to produce low distortion in the model output after applying perturbations to the unselected features in the input while remaining relatively sparse. The authors of [10] also applied RDE to non-canonical input representations to explain model decisions in challenging domains such as audio classification [7] and radio-map estimation [16, 15]. The explanation principle of optimizing a mask $s \in [0,1]^n$ was first proposed by [8] who explained image classification decisions by considering one of the two "deletion games": (1) optimizing for the smallest deletion mask that causes the class score to drop significantly or (2) optimizing for the largest deletion mask that has no significant effect on the class score. The original RDE approach [18] is based on the second deletion game. We decided to work within the RDE framework, due to its flexible mathematical formulation. However, we note that other viable mask-based explanation frameworks such as RISE [23], which does not assume access to the model gradient, exist. Other explanation methods developed by the research community are typically either (1) gradient-based such as Smoothgrad [30], Integrated Gradients [32], and Grad-CAM [26], (2) surrogate models such

as LIME [24], (3) based on propagation of activations in neurons such as LRP [1, 28], and DeepLIFT [28], (4) based on Shapely values from game-theory [17], (6) concept-based such as Concept Activation Vectors [12], or (7) based on generative causal explanations [22]. Also related are methods that were developed to explain individual neurons such as in [21, 6]. To our knowledge, all existing explainability methods operate in pixel space and all methods looking for sparse explanations demand sparsity in pixel space [18, 8, 2].

3 Background: RDE

In this section, we review the rate-distortion explanation (RDE) framework, which was introduced by [18] and later extended by [10] by applying RDE to non-canonical input representations. Suppose $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is a pre-trained model, *e.g.*, a classifier (with m class labels) or a regression model (with m-dimensional output), where n denotes the dimension of the model input. RDE produces an explanation for a model decision $\Phi(x)$ with $x \in \mathbb{R}^n$ as a relatively sparse mask $s \in \{0,1\}^n$ marking the relevant input features in x. More precisely, RDE aims to solve the following constrained optimization problem over a mask $s \in \{0,1\}^n$:

$$\min_{s\in\{0,1\}^n:\, \|s\|_0 \le \ell} \; \mathop{\mathbb{E}}_{v\sim\mathcal{V}} \Big[d\Big(\Phi(x), \Phi(x \odot s + (1-s)\odot v)\Big)\Big] \tag{1}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication), $d(\Phi(x), \cdot)$ is a measure of distortion (*e.g.*, $d(\Phi(x), \cdot) = \|\Phi(x) - \cdot\|_2$), $\mathcal{V}$ is a distribution over input perturbations $v \in \mathbb{R}^n$, and $\ell \in \{1, ..., n\}$ is a given sparsity level for the explanation mask s. A solution s^* to the optimization problem (1) masks relatively few components in the model input x that suffice to approximately retain the model output $\Phi(x)$. This approach is in the spirit of rate-distortion theory, which deals with lossy compression of data. Therefore, [18] coined such explanations *rate-distortion explanations* (RDEs).

In practice, the RDE optimization problem is relaxed to continuous masks $s \in [0,1]^n$ solving:

$$\min_{s\in[0,1]^n} \; \mathop{\mathbb{E}}_{v\sim\mathcal{V}} \Big[d\Big(\Phi(x), \Phi(x \odot s + (1-s)\odot v)\Big)\Big] + \lambda \|s\|_1 \tag{2}$$

In the relaxed optimization problem, the sparsity level of the mask is determined by $\lambda > 0$ and an approximate solution can be found with stochastic gradient descent in $s \in [0,1]^n$ if Φ is differentiable. The authors of [18] applied the RDE method as described above to image classifiers in the pixel domain of images, where each mask entry $s_i \in [0,1]$ corresponds to the i-th pixel values. We refer to this method as *Pixel RDE* throughout this work.

4 RDE Reformulated and Reinterpreted

Instead of applying RDE to the standard input representation $x = [x_1 \dots x_n]^T$, we can apply RDE to a different representation of x to answer a particular

interpretation query. For example, consider a 1D-signal $x \in \mathbb{R}^n$: if we ask "What is the smooth part in the signal x that leads to the model decision $\Phi(x)$?", then we can apply RDE in the Fourier basis of x. Since frequency-sparse signals are smooth, applying RDE in the Fourier basis of x extracts the relevant smooth part of the signal. To accommodate such interpretation queries, we reformulate RDE in Section 4.1. Finally, based on the reformulation, we reinterpret RDE in Section 4.2. Later in Section 5, we use our reformulation and reinterpretation of RDE to derive and motivate CartoonX as a special case and novel explanation method tailored towards image classifiers.

4.1 General Formulation

An input signal $x = [x_1, \dots, x_n]^T$ is represented in a basis $\{b_1, \dots, b_n\}$ as a linear combination $\sum_{i=1}^n h_i b_i$ with coefficients $[h_i]_{i=1}^n$. As we argued above and demonstrate later on, some choices for a basis may be more suitable than others to explain a model decision $\Phi(x)$. Therefore, we define the RDE mask not only on the canonical input representation $[x_i]_{i=1}^n$ but also on a different representation $[h_i]_{i=1}^n$ with respect to a choice of basis $\{b_1, \dots, b_n\}$. Examples of non-canonical choices for a basis include the Fourier basis and the wavelet basis. This work is centered around CartoonX, which applies RDE in the wavelet basis, *i.e.*, a linear data representation. Nevertheless, there also exist other domains and interpretation queries where applying RDE to a non-linear data representation can make sense (see the interpretation query "Is phase or magnitude more important for an audio classifier?" in [10]). Therefore, we formulate RDE in terms of a data representation function $f : \prod_{i=1}^k \mathbb{R}^c \to \mathbb{R}^n,\ f(h_1, \dots, h_k) = x$, which does not need to be linear and allows to mask c channels in the input at once. In the important linear case and $c = 1$, we have $f(h_1, \dots, h_k) = \sum_{i=1}^k h_i b_i$, where $\{b_i, \dots, b_k\} \subset \mathbb{R}^n$ are k fixed vectors that constitute a basis. The case $c > 1$ is useful when one wants to mask out several input channels at once, *e.g.*, all color channels of an image, to reduce the number of entries in the mask that will operate on $[h_i]_{i=1}^k$. In the following, we introduce the important definitions of *obfuscations*, *expected distortion*, *the RDE mask*, and *RDE's ℓ_1-relaxation*, which generalize the RDE framework of [18] to abstract input representations.

Definitions The first two key concepts in RDE are *obfuscations* and *expected distortions*, which are defined below.

Definition 1 (Obfuscations and expected distortions). *Let $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ be a model and $x \in \mathbb{R}^n$ a data point with a data representation $x = f(h_1, ..., h_k)$ as discussed above. For every mask $s \in [0,1]^k$, let $\mathcal{V}$ be a probability distribution over $\prod_{i=1}^k \mathbb{R}^c$. Then the* obfuscation *of x with respect to s and $\mathcal{V}$ is defined as the random vector $y := f(s \odot h + (1-s) \odot v)$, where $v \sim \mathcal{V}$, $(s \odot h)_i = s_i h_i \in \mathbb{R}^c$ and $((1-s) \odot v)_i = (1-s_i)v_i \in \mathbb{R}^c$, for $i \in \{1, \dots, k\}$. A choice for the distribution $\mathcal{V}$ is called* obfuscation strategy. *Furthermore, the* expected distortion *of x with*

respect to the mask s and the perturbation distribution $\mathcal{V}$ is defined as

$$D(x, s, \mathcal{V}, \Phi) := \mathop{\mathbb{E}}_{v\sim\mathcal{V}}\Big[d\Big(\Phi(x), \Phi(y)\Big)\Big],$$

where $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ is a measure of distortion between two model outputs.

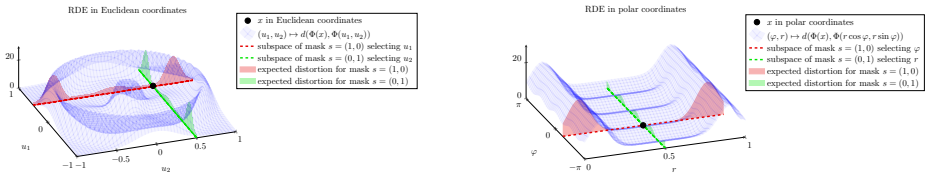
In the RDE framework, the explanation is given by a mask that minimizes distortion while remaining relatively sparse. The rate-distortion explanation mask is defined as follows.

Definition 2 (The RDE mask). *In the setting of Definition 1, we define the* RDE mask *as a solution $s^*(\ell)$ to the minimization problem*

$$\min_{s\in\{0,1\}^k} \quad D(x, s, \mathcal{V}, \Phi) \quad s.t. \quad \|s\|_0 \le \ell, \tag{3}$$

where $\ell \in \{1, \dots, k\}$ is the desired level of sparsity.

Geometrically, the RDE mask s is associated with a particular subspace. The complement mask $(1 - s)$ can be seen as selecting a large stable subspace of Φ, where each point represents a possible perturbation in unselected coefficients in h. The RDE mask minimizes the expected distortion along its associated subspace, which requires non-local information of Φ. We illustrate this geometric view of RDE in Figure 1 with a toy example for a hypothetical classifier $\Phi : \mathbb{R}^2 \to \mathbb{R}^m$ and two distinct input representations: (1) Euclidean coordinates, *i.e.*, f is the identity in $x = f(h)$, and (2) polar coordinates, *i.e.*, $f(h) = (h_2 \cos h_1, h_2 \sin h_1) = x$. In the example, we assume $\mathcal{V}$ to be a uniform



(a) (b)

Fig. 1: RDE for a hypothetical toy-example in (a) Euclidean coordinates and (b) polar coordinates. Here, the RDE mask can find low expected distortion in polar coordinates but not in Euclidean coordinates. Therefore, in this example, polar coordinates are more appropriate to explain $\Phi(x)$, and RDE would determine that the angle φ, not the magnitude r, is relevant for $\Phi(x)$.

distribution on $[-1, 1]^2$ in the Euclidean representation and a uniform distribution on $[-\pi, \pi] \times [0, 1]$ in the polar representation. The expected distortion associated with the masks $s = (1, 0)$ and $s = (0, 1)$ is given by the red and green shaded area, respectively. The RDE mask aims for low expected distortion, and

hence, in polar coordinates, the RDE mask would be the green subspace, *i.e.*, $s = (0, 1)$. On the other hand, in Euclidean coordinates, neither $s = (1, 0)$ nor $s = (0, 1)$ produces a particularly low expected distortion, making the Euclidean explanation less meaningful than the polar explanation. The example illustrates why certain input representations can yield more meaningful explanatory insight for a given classifier than others—an insight that underpins our novel CartoonX method. Moreover, the plot in polar coordinates illustrates why the RDE mask cannot be simply chosen with local distortion information, *e.g.*, with the lowest eigenvalue of the Hessian of $h \mapsto d(\Phi(x), \Phi(f(h)))$: the lowest eigenvalue in polar coordinates belongs to the red subspace and does not see the large distortion on the tails.

As was shown by [18], the RDE mask from Definition 2 cannot be computed efficiently for non-trivial input sizes. Nevertheless, one can find an approximate solution by considering continuous masks $s \in [0, 1]^k$ and encouraging sparsity through the ℓ_1-norm.

Definition 3 (RDE's ℓ_1-relaxation). *In the setting of Definition 1, we define* RDE's ℓ_1-relaxation *as a solution $s^*(\lambda)$ to the minimization problem*

$$\min_{s \in [0,1]^k} \quad D(x, s, \mathcal{V}, \Phi) + \lambda \|s\|_1, \tag{4}$$

where $\lambda > 0$ is a hyperparameter for the sparsity level.

The ℓ_1-relaxation above can be solved with stochastic gradient descent (SGD) over the mask s while approximating $D(x, s, \mathcal{V}, \Phi)$ with i.i.d. samples from $v \sim \mathcal{V}$.

Obfuscation Strategies An obfuscation strategy is defined by the choice of the perturbation distribution $\mathcal{V}$. Common choices are Gaussian noise [18, 8], blurring [8], constants [8], and inpainting GANs [10, 2]. Inpainting GANs train a generator $G(s, z, h)$ (z denotes random latent factors) such that for samples $v \sim G(s, z, h)$ the obfuscation $f(s \odot h + (1 - s) \odot v)$ remains in the data manifold. In our work, we refrain from using an inpainting GAN due to the following reason: it is hard to tell whether a GAN-based mask did not select coefficients because they are unimportant or because the GAN can easily inpaint them from a biased context (*e.g.*, a GAN that always inpaints a car when the mask shows a traffic light). We want to explain a black-box method transparently, which is why we opt for a simple distribution on the price of not accurately representing the data distribution. We choose a simple and well-understood obfuscation strategy, which we call *Gaussian adaptive noise*. It works as follows: Let $A_1, ..., A_j$ be a pre-defined choice for a partition of $\{1, \dots, k\}$. For $i = 1, ..., j$, we compute the empirical mean and empirical standard deviation for each A_i:

$$\mu_i \coloneqq \frac{\sum_{a \in A_i, t=1,\dots,d_a} h_{at}}{\sum_{a \in A_i} d_a}, \qquad \sigma_i \coloneqq \sqrt{\frac{1}{\sum_{a \in A_i} d_a} \sum_{a \in A_i, t=1,\dots,d_a} (\mu_i - h_{at})^2} \tag{5}$$

The adaptive Gaussian noise strategy then samples $v_{at} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for all members $a \in A_i$ and channels $t = 1, ..., d_a$. We write $v \sim \mathcal{N}(\mu, \sigma^2)$ for the resulting

Gaussian random vector $v \in \prod_{i=1}^{k} \mathbb{R}^c$. For Pixel RDE, we only use one set $A_1 = \{1, ..., k\}$ for all k pixels. In CartoonX, which represents input signals in the discrete wavelet domain, we partition $\{1, ..., k\}$ along the scales of the discrete wavelet transform.

Measures of distortion There are various choices for the measure of distortion $d(\Phi(x), \Phi(y))$. For example, one can take the squared distance in the post-softmax probability of the predicted label for x, *i.e.*, $d\big(\Phi(x), \Phi(y)\big) := \big(\Phi_{j^*}(x) - \Phi_{j^*}(y)\big)^2$, where $j^* := \arg\max_{i=1,...,m} \Phi_i(x)$ and $\Phi(x)$ is assumed to be the post-softmax probabilities of a neural net. Alternatively, one could also choose $d(\Phi(x), \Phi(y))$ as the ℓ_2-distance or the KL-Divergence in the post-softmax layer of Φ. In our experiments for CartoonX, we found that these choices had no significant effect on the explanation (see Figure 8d).

4.2 Interpretation

The philosophy of the generalized RDE framework is that an explanation for a decision $\Phi(x)$ on a generic input signal $x = f(h)$ should be some simplified version of the signal, which is interpretable to humans. The simplification is achieved by demanding sparsity in a suitable representation system h, *which sparsely represents the class of explanations that are desirable for the interpretation query*. This philosophy is the fundamental premise of CartoonX, which aims to answer the interpretation query *"What is the relevant piece-wise smooth part of the image for a given image classifier?"*. CartoonX first employs RDE on a representation system $x = f(h)$ that sparsely represents piece-wise smooth images and finally visualizes the relevant piece-wise smooth part as an image back in pixel space. In the following section, we explain why wavelets provide a suitable representation system in CartoonX, discuss the CartoonX implementation, and evaluate CartoonX qualitatively and quantitatively on ImageNet.

5 CartoonX

The focus of this paper is *CartoonX*, a novel explanation method—tailored to image classifications—that we obtain as a special case of our generalized RDE framework formulated in Section 4. CartoonX first performs RDE in the discrete wavelet position-scale domain of an image x, and finally, visualizes the wavelet mask s as a piece-wise smooth image in pixel space. Wavelets provide optimal representations for piece-wise smooth 1D functions [5], and represent 2D piece-wise smooth images, also called *cartoon-like images* [14], efficiently as well [25]. In particular, sparse vectors in the wavelet coefficient space encode cartoon-like images reasonably well [19]—certainly better than sparse pixel representations. Moreover, wavelets constitute an established tool in image processing [20].

The optimization process underlying CartoonX produces sparse vectors in the discrete wavelet coefficient space, which results in cartoon-like images as

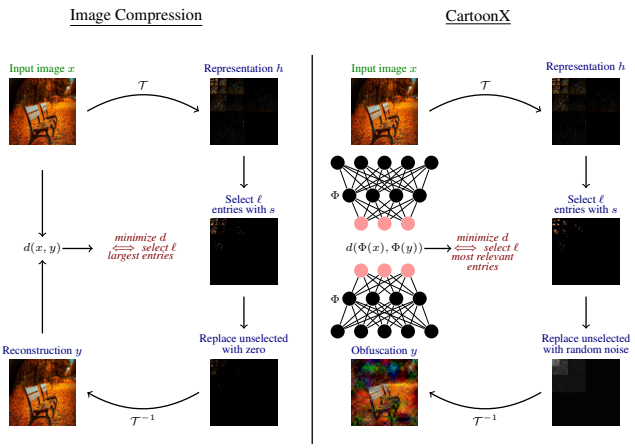


Fig. 2: CartoonX shares many interesting parallels to wavelet-based image compression. Distortion is denoted as d, Φ is an image classifier, h denotes the discrete wavelet coefficients, $\mathcal{T}$ is the discrete wavelet transform, and ℓ is the coefficient budget.

explanations. This is the fundamental difference to Pixel RDE, which produces rough, jittery, and pixel-sparse explanations. Cartoon-like images provide a natural model of simplified images. Since the goal of the RDE framework is to generate an easy to interpret simplified version of the input signal, we argue that CartoonX explanations are more appropriate for image classification than Pixel RDEs. Previous work, such as Grad-CAM [26], produces smooth explanations, which also avoid jittery explanations. CartoonX produces *roughly piece-wise smooth explanations and not smooth explanations*, which we believe to be more appropriate for images, since smooth explanations cannot preserve edges well. Moreover, we believe that CartoonX enforces piece-wise smoothness in a mathematically more natural manner than explicit smoothness regularization (as in [9]) because wavelets sparsely represent piece-wise smooth signals well. Therefore, CartoonX does not rely on additional smoothness hyperparameters.

CartoonX exhibits interesting parallels to wavelet-based image compression. In image compression, distortion is minimized in the image domain, which is equivalent to selecting the ℓ largest entries in the discrete wavelet transform (DWT) coefficients. CartoonX minimizes distortion in the model output of Φ, which translates to selecting the ℓ most relevant entries in the DWT coefficients. The objective in image compression is efficient data representation, *i.e.*, producing minimal data distortion with a budget of ℓ entries in the DWT coefficients. Conversely, in CartoonX, the objective is extracting the relevant piece-wise smooth part, *i.e.*, producing minimal model distortion with a budget of ℓ entries in the DWT coefficients. We illustrate this connection in Figure 2—highlighting once more the *rate-distortion* spirit of the RDE framework.

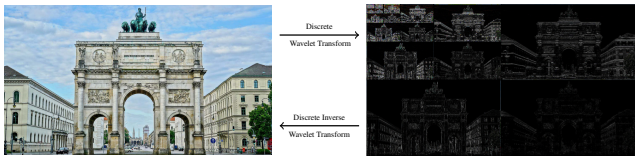


Fig. 3: Visualization of the DWT coefficients for five scales. Three L-shaped sub-images describe coefficients for details in vertical, horizontal, and diagonal orientation at a particular scale. The largest sub-images (the outer L-shape) belong to the lowest scale, *i.e.*, the highest resolution. The smaller L-shaped sub-images gradually build up to higher scales, *i.e.*, lower resolution features.

5.1 Implementation

An image $x \in [0,1]^n$ with $c \in \{1,3\}$ channels, $k \in \mathbb{N}$ pixels can be represented in a wavelet basis by computing its DWT, which is defined by the number of scales $J \in \{1,\dots,\lfloor \log_2 k \rfloor\}$, the padding mode, and a choice of the mother wavelet (*e.g.*, Haar or Daubechies). For images, the DWT computes four types of coefficients: details in (1) horizontal, (2) vertical, and (3) diagonal orientation at scale $j \in \{1,\dots,J\}$, and (4) coefficients of the image at the very coarsest resolution. We briefly illustrate the DWT for an example image in Figure 3.

CartoonX, as described in Algorithm 1, computes the RDE mask in the wavelet domain of images. More precisely, for the data representation $x = f(h)$, we choose h as the concatenation of all the DWT coefficients along the channels, *i.e.*, $h_i \in \mathbb{R}^c$. The representation function f is then the discrete inverse wavelet transform, *i.e.*, the summation of the DWT coefficients times the DWT basis vectors. We optimize the mask $s \in [0,1]^k$ on the DWT coefficients $[h_1,\dots,h_k]^T$ to minimize RDE's ℓ_1-relaxation from Definition 3. For the obfuscation strategy $\mathcal{V}$, we use adaptive Gaussian noise with a partition by the DWT scale (see Section 4.1), *i.e.*, we compute the empirical mean and standard deviation per scale. To visualize the final DWT mask s as a piece-wise smooth image in pixel space, we multiply the mask with the DWT coefficients of the greyscale image $\hat{x}$ of x before inverting the product back to pixel space with the inverse DWT. The pixel values of the inversion are finally clipped into $[0,1]$ as are obfuscations during the RDE optimization to avoid overflow (we assume here the pixel values in x are normalized into $[0,1]$). The clipped inversion in pixel space is the final CartoonX explanation.

5.2 Experiments

We compare CartoonX to the closely related Pixel RDE [18] and several other state-of-the-art explanation methods, *i.e.*, Integrated Gradients [32], Smoothgrad [30], Guided Backprop [31], LRP [1], Guided Grad-CAM [27], Grad-CAM [27], and LIME [24]. Our experiments use the pre-trained ImageNet classifiers MobileNetV3-Small [11] (67.668% top-1 acc.) and VGG16 [29] (71.592% top-1

Algorithm 1: CartoonX

Data: Image $x \in [0,1]^n$ with c channels and k pixels, pre-trained classifier Φ.
Initialization: Initialize mask $s \coloneqq [1,...,1]^T$ on
DWT coefficients $h = [h_1,...,h_k]^T$ with $x = f(h)$, where f is the inverse DWT.
Choose sparsity level $\lambda > 0$, number of steps N, number of noise samples L, and measure of distortion d.
for $i \leftarrow 1$ **to** N **do**
  Sample L adaptive Gaussian noise samples $v^{(1)},...,v^{(L)} \sim \mathcal{N}(\mu, \sigma^2)$;
  Compute obfuscations $y^{(1)},...,y^{(L)}$ with $y^{(i)} \coloneqq f(h \odot s + (1-s) \odot v^{(i)})$;
  Clip obfuscations into $[0,1]^n$;
  Approximate expected distortion $\hat{D}(x,s,\Phi) \coloneqq \sum_{i=1}^{L} d(\Phi(x), \Phi(y^{(i)}))^2/L$;
  Compute loss for the mask, *i.e.*, $\ell(s) \coloneqq \hat{D}(x,s,\Phi) + \lambda\|s\|_1$;
  Update mask s with gradient descent step using $\nabla_s \ell(s)$ and clip s back to $[0,1]^k$;
end
Get DWT coefficients $\hat{h}$ for greyscale image $\hat{x}$ of x;
Set $\mathcal{E} \coloneqq f(\hat{h} \odot s)$ and finally clip $\mathcal{E}$ into $[0,1]^k$;

acc.). Images were preprocessed to have 256×256 pixel values in $[0,1]$. Throughout our experiments with CartoonX and Pixel RDE, we used the Adam optimizer [13], a learning rate of $\epsilon = 0.001$, $L = 64$ adaptive Gaussian noise samples, and $N = 2000$ steps. Several different sparsity levels were used. We specify the sparsity level in terms of the number of mask entries k, *i.e.*, by choosing the product λk. Pixel RDE typically requires a smaller sparsity level than CartoonX. We chose $\lambda k \in [20, 80]$ for CartoonX and $\lambda k \in [3, 20]$ for Pixel RDE. The obfuscation strategy for Pixel RDE was chosen as Gaussian adaptive noise with mean and standard deviation computed for all pixel values (see Section 4.1). We implemented the DWT for CartoonX with the Pytorch Wavelets package, which is compatible with PyTorch gradient computations, and chose the Daubechies 3 wavelet system with $J = 5$ scales and zero-padding. For the Integrated Gradients method, we used 100 steps, and for the Smoothgrad method, we used 10 samples and a standard deviation of 0.1.

Interpreting CartoonX In order to correctly interpret CartoonX, we briefly review important properties of the DWT. To cover a large area in an image with a constant value or slowly and smoothly changing gray levels, it suffices to select very few high-scale wavelet coefficients. Hence, for the wavelet mask in CartoonX, it is cheap to cover large image regions with constant or blurry values. Conversely, one needs many high-scale wavelet coefficients to produce fine details such as edges in an image, so fine details are expensive for CartoonX. Hence, the fine details present in the CartoonX are important features for the outcome of the classifier, and fine image features that are replaced by smooth areas in CartoonX are not important for the classifier. It is important to keep in mind that the final CartoonX explanation is a visualization of the wavelet mask

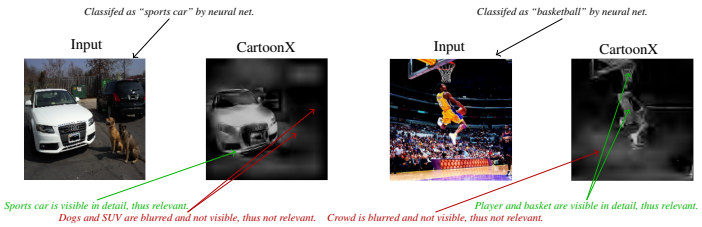
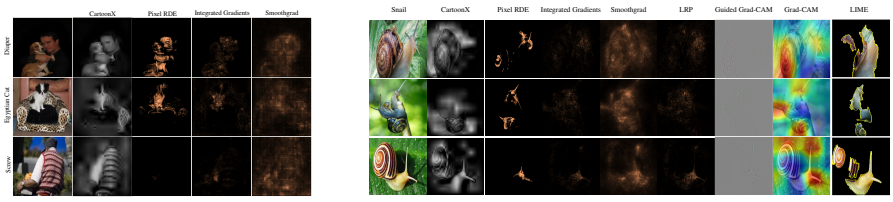


Fig. 4: The CartoonX explanation is an image that suffices to retain the classification decision. For the sports car, CartoonX blurs out the SUV and the dogs. This means the dog and the SUV are irrelevant. For the basketball, the crowd is blurred out. This means the crowd is not relevant since the player and the basket with the crowd blurred out retains the classification as "basketball". The left example also shows that CartoonX is class-discriminative since it blurs out the dogs and the SUV, which belong to other classes.

in pixel space, and *should not be interpreted as a pixel-mask or ordinal pixel-attribution.* CartoonX is not a saliency-map or heatmap but an explanation that is to be interpreted as an image that suffices to retain the classification decision. We illustrate this point in Figure 4 with two examples.



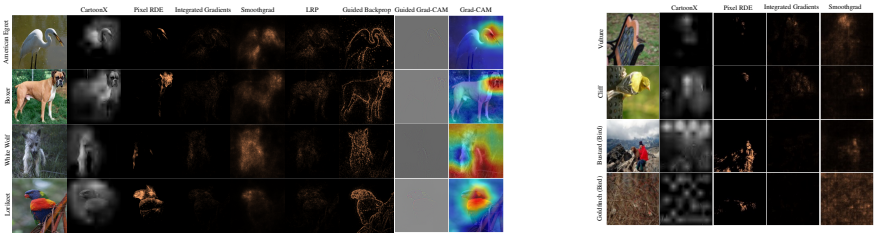
(a) (b)

Fig. 5: (a) Each row compares CartoonX explanations of misclassifications by MobileNetV3-Small. The predicted label is depicted next to each misclassified image. (b) Comparing CartoonX explanations for VGG16 for three different images of correctly classified snails.

Qualitative Evaluation In practice, explaining misclassifications is particularly relevant since good explanations can pinpoint model biases and causes for model failures. In Figure 5a, we illustrate how CartoonX can help explain misclassified examples by revealing classifier-relevant piece-wise smooth patterns that are not easily visible in other pixel-based methods. In the first row in Figure 5a, the input image shows a man holding a dog that was classified as a "diaper". CartoonX shows the man not holding a dog but a baby, possibly revealing that the neural net associated diapers with babies and babies with the pose with

which the man is holding the dog. In the second row, the input image shows a dog sitting on a chair with leopard patterns. The image was classified as an "Egyptian Cat", which can exhibit leopard-like patterns. CartoonX exposes the Egyptian cat by connecting the dog's head to parts of the armchair forming a cat's torso and legs. In the last row, the input image displays the backside of a man wearing a striped sweater that was classified as a "screw". CartoonX reveals how the stripe patterns look like a screw to the neural net.

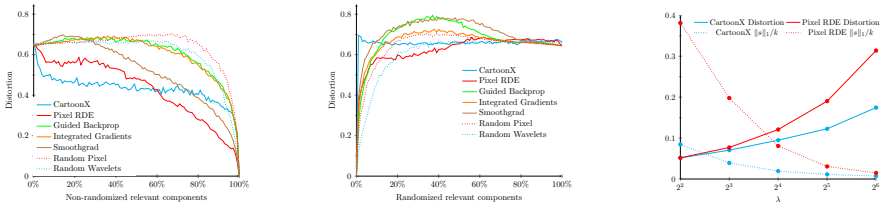
Figure 5b further compares CartoonX explanations of correct classifications by VGG16. We also compare CartoonX on random ImageNet samples in Figure 6a to provide maximal transparency and fair qualitative comparison. In Figure 6b, we also show failures of CartoonX. These are examples of explanations that are not interpretable and seem to fail at explaining the model prediction. Notably, most failure examples are also not particularly well explained by other state-of-the-art methods. It is challenging to state with certainty the underlying reason for the CartoonX failures since there it is always possible that the neural net bases its decision on non-interpretable grounds.



(a) CartoonX on random samples. (b) Examples of CartoonX failures.

Fig. 6: On random Imagenet samples, CartoonX consistently produces interpretable explanations. Established explanation methods tend to also be difficult to interpret on CartoonX's failure examples.

Quantitative Evaluation To compare CartoonX quantitatively against other explanation methods, we computed explanations for 100 random ImageNet samples and ordered the image coefficients (for CartoonX the wavelet coefficients) by their respective relevance score. Figure 7a plots the rate-distortion curve, *i.e.*, the distortion achieved in the model output (measured as the ℓ_2-norm in the post-softmax layer) when keeping the most relevant coefficients and randomizing the others. We expect a good explanation to have the most rapid decaying rate-distortion curve for low rates (non-randomized components), which is the case for CartoonX. Note that the random baseline in the wavelet representation is not inherently more efficient than the random baseline in the pixel representation. Moreover, Figure 7b plots the achieved distortion versus the fraction of randomized relevant components. Here, we expect a good explanation to have the sharpest early increase, which CartoonX again realizes. Lastly, Figure 7c



(a) (b) (c)

Fig. 7: In (a) the best explanation exhibits steepest early decay. In (b) best explanation exhibits sharpest early increase. In (c) best explanation exhibits lowest distortion and lowest normalized ℓ_1-norm of mask (*i.e.*, highest sparsity).

plots the distortion and non-sparsity (measured as the normalized ℓ_1-norm) of the RDE mask for Pixel RDE and CartoonX at different λ values. The plot underscores the efficiency advantage of CartoonX over Pixel RDE since CartoonX achieves lower distortion and higher sparsity throughout all λ values. For all three plots, random perturbations were drawn from the adaptive Gaussian distribution described in Section 4.1.

Sensitivity to Hyperparameters We compare qualitatively CartoonX's sensitivity to its primary hyperparameters. Figure 8a plots CartoonX explanations and Pixel RDEs for increasing λ. We conistently find that CartoonX is less sensitive than Pixel RDE to λ. In practice, this means one can find a suitable λ faster for CartoonX than for Pixel RDE. Note that for $\lambda = 0$, Pixel RDE is entirely yellow because the mask is initialized as $s = [1 \dots 1]^T$ and $\lambda = 0$ provides no incentive to make s sparser. For the same reason, CartoonX is simply the greyscale image when $\lambda = 0$. Figure 8b plots CartoonX explanations for two choices of $\mathcal{V}$: (1) Gaussian adaptive noise (see Section 4.1) and (2) constant zero perturbations. We observe that the Gaussian adaptive noise gives much more meaningful explanations than the simple zero baseline perturbations. Figure 8d plots CartoonX explanations for four choices of $d(\Phi(x), \Phi(y))$, where x is the original input, y is the RDE obfuscation, and Φ outputs post-softmax probabilities: (1) squared ℓ_2 in probability of predicted label j^*, (2) $d(\Phi(x), \cdot) = \|\Phi_{j^*}(x) - 1\|$, *i.e.*, distance that maximizes probability of predicted label, (3) ℓ_2 in post-softmax, (4) KL-Divergence in post-softmax. We do not observe a significant effect by the distortion measure on the explanation. Finally, in Figure 8c we compare the effect of the mother wavelet in the DWT on the CartoonX explanation. All choices of mother wavelets (labeled as in the Pytorch Wavelets package) provide consistent explanations except for the Haar wavelet, which produces images built of large square pixels.

Limitations For MobileNetV3-Small, an image of 256×256 pixels, 16 noise samples, and 2000 optimization steps, we reported a runtime of 45.10s for CartoonX and 34.09s for Pixel RDE on the NVIDIA Titan RTX GPU. CartoonX

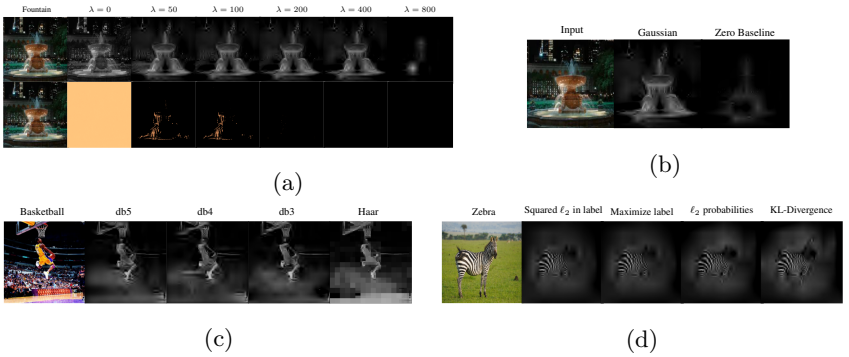


Fig. 8: (a) Top row depicts CartoonX, and the bottom row depicts Pixel RDE, for increasing values of λ. CartoonX for different (b) perturbation distributions, (c) mother wavelets, (d) distortion measures.

is only slightly slower than Pixel RDE. However, like other perturbation-based methods, CartoonX is significantly slower than gradient or propagation-based methods, which only compute a single or few forward and backward passes and are very fast (Integrated Gradients computes an explanation in 0.48s for the same image, model, and hardware). We acknowledge that the runtime for CartoonX in its current form constitutes a considerable limitation for many critical applications. However, we are confident that we can significantly reduce the runtime in future work by either learning a strong initial wavelet mask with a neural net or even learning the final wavelet mask with a neural net, similar to the real time image saliency work in [3]. Finally, solving RDE's ℓ_1-relaxation requires access to the model's gradients. Hence, CartoonX is limited to differentiable models.

6 Conclusion

CartoonX is the first explanation method for differentiable image classifiers based on wavelets. We corroborated experimentally that CartoonX can reveal novel explanatory insight and achieves a better rate-distortion than state-of-the-art methods. Nonetheless, CartoonX is still computationally quite expensive, like other perturbation-based explanation methods. In the future, we hope to devise new techniques to speed up the runtime for CartoonX and study the effect of using inpainting GANs for perturbations. We believe CartoonX is a valuable new explanation method for practitioners and potentially a great source of inspiration for future explanation methods tailored to specific data domains.

Acknowledgments GK was supported in part by the ONE Munich Strategy Forum as well as by Grant DFG-SFB/TR 109, Project C09 and DFG-SPP-2298, KU 1446/31-1 and KU 1446/32-1.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
2. Chang, C., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: Proceedings of the 7th International Conference on Learning Representations, ICLR (2019)
3. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: NIPS (2017)
4. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey. ArXiv **abs/2006.11371** (2020)
5. DeVore, R.A.: Nonlinear approximation. Acta Numerica **7**, 51–150 (1998)
6. Dhamdhere, K., Sundararajan, M., Yan, Q.: How important is a neuron. In: International Conference on Learning Representations (2019)
7. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., Simonyan, K.: Neural audio synthesis of musical notes with wavenet autoencoders. In: Proceedings of the 34th International Conference on Machine Learning, ICML. vol. 70, p. 1068–1077 (2017)
8. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3449–3457 (2017)
9. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
10. Heiß, C., Levie, R., Resnick, C., Kutyniok, G., Bruna, J.: In-distribution interpretability for challenging modalities. Preprint arXiv:2007.00758 (2020)
11. Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q.: Searching for MobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019)
12. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: ICML (2018)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
14. Kutyniok, G., Lim, W.Q.: Compactly supported shearlets are optimally sparse. Journal of Approximation Theory **163**(11), 1564–1589 (2011)
15. Levie, R., Yapar, C., Kutyniok, G., Caire, G.: Pathloss prediction using deep learning with applications to cellular optimization and efficient d2d link scheduling. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8678–8682 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053347
16. Levie, R., Yapar, C., Kutyniok, G., Caire, G.: RadioUNet: Fast radio map estimation with convolutional neural networks. IEEE Transactions on Wireless Communications **20**(6), 4001–4015 (2021)
17. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NeurIPS. p. 4768–4777 (2017)

18. Macdonald, J., Wäldchen, S., Hauch, S., Kutyniok, G.: A rate-distortion framework for explaining neural network decisions. Preprint arXiv:1905.11092 (2019)
19. Mallat, S.: A Wavelet Tour of Signal Processing (Third Edition), chap. 11.3. Academic Press, third edition edn. (2009)
20. Mallat, S.: A Wavelet Tour of Signal Processing (Third Edition). Academic Press, Boston, third edition edn. (2009)
21. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems (NIPS) (2016)
22. O'Shaughnessy, M., Canal, G., Connor, M., Rozell, C., Davenport, M.: Generative causal explanations of black-box classifiers. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 5453–5467. Curran Associates, Inc. (2020)
23. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: BMVC (2018)
24. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD. p. 1135–1144. Association for Computing Machinery (2016)
25. Romberg, J.K., Wakin, M.B., Baraniuk, R.G.: Wavelet-domain approximation and compression of piecewise smooth images. IEEE Trans. Image Processing **15**, 1071–1087 (2006)
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
27. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**, 336–359 (2019)
28. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning, ICML. vol. 70, p. 3145–3153 (2017)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
30. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. In: Workshop on Visualization for Deep Learning, ICML (2017)
31. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
32. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML. vol. 70, p. 3319–3328 (2017)