Sampling Based On Natural Image Statistics Improves Local Surrogate Explainers

Ricardo Kleinlein[1]
ricardo.kleinlein@upm.es
Alexander Hepburn[2]
alex.hepburn@bristol.ac.uk
Raúl Santos-Rodríguez[2]
enrsr@bristol.ac.uk
Fernando Fernández-Martínez[1]
fernando.fernandezm@upm.es

[1] Information Processing and Telecommunications Center
E.T.S.I. de Telecomunicación,
Universidad Politécnica de Madrid
Madrid, Spain

[2] Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom

Abstract

Many problems in computer vision are recently been tackled using deep neural networks, whose predictions cannot be easily interpreted. Surrogate explainers aim to address this, as a popular post-hoc interpretability method to further understand how a black-box model arrives at a particular prediction. By training a simple, more interpretable model to locally approximate the decision boundary of a non-interpretable system, we can estimate the relative importance of the input features on the prediction. Focusing on images, most surrogate explainers, e.g., LIME, generate a local neighbourhood around a query image by sampling in an interpretable domain. However, interpretable domains have traditionally been derived exclusively from the intrinsic features of the query image, not taking into consideration the manifold of the data the non-interpretable model has been exposed to in training (or more generally, the manifold of real images). This leads to suboptimal surrogates as they are trained on images that lie within low probability regions of the manifold of real images. In this work, we address this limitation by aligning the local neighbourhood on which the surrogate is trained with the original training data distribution, even when this distribution is not accessible. We propose two approaches to do so, namely (1) altering the method for sampling the local neighbourhood and (2) using perceptual metrics to convey some of the properties of the statistics of natural images.

1 Introduction

Deep neural networks are at the forefront of both computer vision research and related industrial applications [2, 8]. Practitioners in the field often use these large uninterpretable models due to their capacity to make accurate predictions, neglecting the ability to understand the rationale behind a prediction. Therefore if a particular prediction does not match our expectations, we lack the resources to assess the reasoning behind it. This undermines the confidence of the users in the whole system [17].

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

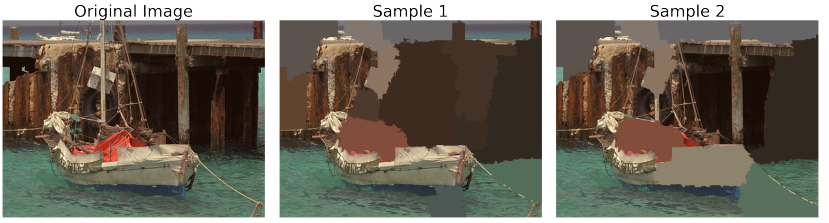
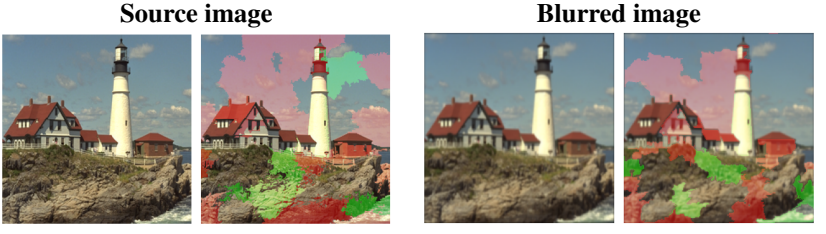


Figure 1: The original image (left) and examples of sampled data using the LIME approach, with mean colour occlusion.



(a) Default LIME: cosine distance & mean color occlusion

(b) Explanations using MS-SSIM & Gaussian blurring

Figure 2: Explanations for the prediction "lighthouse" on source and blurred image. Green and red regions denote superpixels supporting or inhibiting the prediction. Our proposal (b) not only makes explanations more similar to what humans would estimate important, but also enhances the robustness of the explanations, as opposed to a canonical LIME (a).

The exact definition of what constitutes an explanation is still a matter of debate [12] and it can be argued that the models themselves should be inherently interpretable [24]. For image classification, "explaining a prediction" generally refers to presenting visual cues that allow users to build an intuition on features of the input that drive the decision-making process of a model, regardless of the accuracy of the prediction. Post-hoc interpretability techniques range from counterfactuals – finding the closest data point of the opposite class [30] – to permuting values of a feature to check the effect this has on a classification prediction [7]. Alternatively, surrogate explainers involve locally approximating the decision boundary around a query point, using a simpler interpretable model [1, 23].

In this work, we focus on surrogate explainers for image classification. In this context, images are usually represented as a collection of superpixels that encapsulate adjacent pixel regions with similar visual properties. In order to build an interpretable domain and train the corresponding local surrogate, a neighbourhood is sampled around the query image. As seen in Fig. 1, the generation techniques used in the most popular surrogate explainer method (LIME) creates samples that are far from what we can consider real images [14, 15].

We argue that this process is sub-optimal and leads to inconsistent explanations. We suggest that the more this neighbourhood resembles the distribution that the global model was trained using (or at least the distribution of natural images), the more robust the explanations will be. We propose two ways of achieving this; using a different sampling method to better exploit the training distribution (if available) or using perceptual metrics that have been shown to encode fundamental properties of the distribution of natural images. In general, when considering images, access to the actual distribution is intractable. It has been shown that perceptual metrics are correlated with the probability of natural images [10], acting as proxy to a space that is more aligned with the distribution of natural images than the Euclidean space. Figure 2 illustrates how explanations built this way compare against those computed from a canonical LIME perspective.

2 Surrogate Explainers

Although it is common practise to evaluate the adequacy of a machine learning model according to a fixed set of metrics like accuracy rates for classification, or mean squared error for regression tasks, these are not sufficient to fully characterise the behaviour of the model. Because of this limitation, a complete understanding of an automatic system should also include the ability to explain its predictions [13]. Arguably, the most popular approach to provide explanations for black-box models are surrogate-based techniques. These are post-hoc local approximations to the decision boundary of a black-box system around a query point x, learned by a simple model that is, in many cases, linear. We define an explanation as

$$\exp_x = \arg\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

where $\mathcal{L}(f, g, \pi_x)$ is the fit of the surrogate model g from the family of surrogates G, f denotes the black-box model and π_x is the neighbourhood sampled around a query data point $x \in \mathcal{X}$. $\Omega(g)$ is a penalty on the complexity of model g. If $\Omega(g)$ is the L2 norm of the weights g, the surrogate model becomes ridge regression. This formulation conveys the idea that our surrogate g should find the best fit to the local boundary decision of the black-box f given the restrictions in complexity expressed by $\Omega(g)$, only around the neighbourhood π_x. Surrogate explainers can be broken up into three interoperable modules; an interpretable data representation, a data sampling procedure and the explanation generation [28]. These modules are required to be carefully chosen by the practitioner in order to address the problem at hand.

Interpretable data representation Samples in the original domain $\mathcal{X}$ are transformed into a human interpretable representation, $\mathcal{Z}$. In the case of natural images, superpixels offer an interpretable domain in which an image can be represented as a binary vector that encodes whether a specific region has been occluded (removed) or altered in any way.

Sampling Data sampling refers to the generation of the neighbourhood π_x. For images, this usually involves sampling binary vectors in the interpretable data representation defined above (see Fig. 1) and replacing the pixels within the occluded superpixel with the mean colour of the superpixel. This process is called mean colour occlusion.

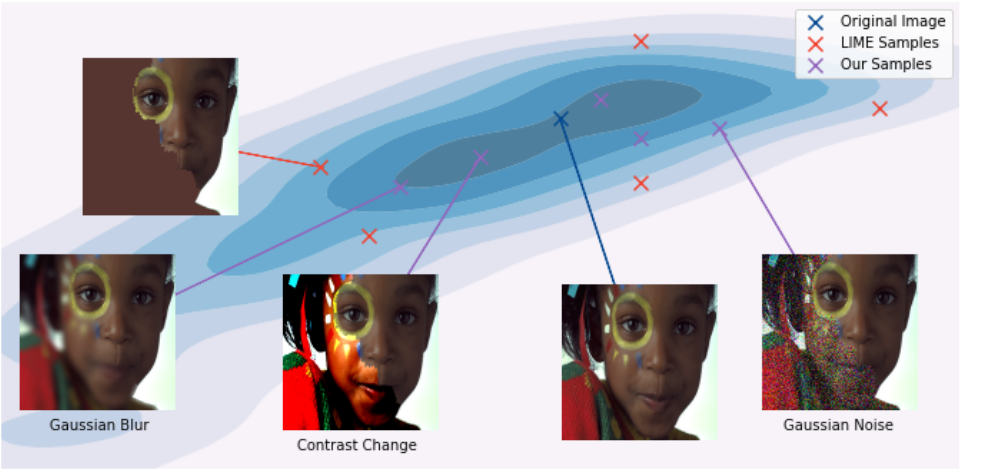


Figure 3: Illustration of the location of our samples and samples according to LIME on an idealised distribution of natural images. Our samples are intended to be drawn from higher probability regions, whereas samples using mean colour occlusion lie in low probability regions.

Explanation generation The final stage is training the surrogate model in order to replicate the behaviour of the black-box model. The surrogate model aims to learn the mapping between instances in π_x sampled from the interpretable domain $z \in \mathcal{Z}$ and the probabilities estimated for a given class by the black-box model, $f(z)$. Consequently, we define a locally weighted square loss:

$$\mathcal{L}(f,g,\pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(x,z)(f(z) - g(z'))^2 \tag{2}$$

3 Surrogates and Natural Image Statistics

Even though it is clear that the choice of π_x is critical when producing reliable explanations,there are limited studies that have investigated this topic [6, 9, 19]. We believe that in order to faithfully mimic the vicinity around a query x to be explained, π_x should incorporate information about the statistics of the distribution the black-box model was trained on. A possible way to do so is to change the sampling process in the interpretable domain $\mathcal{Z}$.

3.1 Sampling according to the statistics of natural images

Sampling is usually performed in the interpretable domain $\mathcal{Z}$. This representation usually has a one-to-one correspondence with the original, non-interpretable domain $\mathcal{X}$ since, in order to train a surrogate, the output of the black-box model $f(\pi_x)$ is required. For tabular data, LIME suggests using a one-to-many mapping and performs inverse sampling, where a Gaussian is fit to the data in one area of $\mathcal{X}$, corresponding to one integer value in the interpretable domain $\mathcal{Z}$. However, this leads to increased randomness and variability between the surrogates produced [6, 28].

Applying LIME to images, the domain of superpixels can be sampled as binary vectors z' from a discrete uniform distribution that results in images z with superpixels ablated according to the binary feature in the sampled vectors. Usually, this binary representation encodes either whether the superpixel is present or not. If 0, all pixel values within the ablated superpixel are set to 0 (zero patching), or set to the mean value within the superpixel (mean colour occlusion). The core idea being if I remove the information in this superpixel, what effect will it have on the prediction. We propose to use more realistic sampling methods in order to capture the statistics of real-world images.

Given direct access to the actual distribution of real-world images, we could replace the ablated superpixel with samples drawn from such distribution. As this is usually intractable, we propose to approximate the effect by transforming the pixels according to distortions often found in natural images. As illustrated in Figure 3, we expect sampling from the proposed distributions leads to more realistic images, with the neighbourhood π_x being closer to the actual distribution used to train the the global model. Alternatively, we can resort to using perceptual metrics that naturally encapsulate properties of the distribution of real images.

3.2 Perceptual metrics

Another fundamental aspect when sampling a neighbourhood π_x is to measure the distance between the generated samples and the query x. A neighbourhood is defined by

$$\pi_x(x,z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \tag{3}$$

where D is a distance between samples z and x and σ is the width of the exponential kernel. The distance used in the original implementation of LIME is the cosine similarity between a vector of all ones representing the superpixels of the original image x' and the binary mask of superpixels in sample z'. However, as image explanations come in the form of visual cues, the explanation we produce must ultimately resemble the way we humans interpret and perceive visual information. As a consequence, the cosine distance may not be the best choice in all scenarios.

This is particularly important for natural images, where properties related to the psychophysics of human perception take a predominant role in explaining a prediction. Perceptual metrics attempt to recreate human psychophysical results when observing a reference and distorted image. Models based on the human visual system have been proved to be effective at this task [11, 16, 31]. One such metric is structural similarity and its multi-scale variant, the *multi-scale structural similarity* index (MS-SSIM) [31]. MS-SSIM is based on the principle that the perceived structural similarity will be preserved despite the distortion. It has been shown that this distance faithfully reflects the perceptual similarity between two images as perceived by the human visual system[10]. In our empirical analysis on natural images we will use MS-SSIM as $D(x,z)$ as an alternative to the cosine distance in order to implicitly alter the distribution of sampled images, making it closer to that of the training data of the black-box model.

4 Experiments

The visual domain poses a particularly challenging environment since sampling from the actual data distribution of all the real images is far from trivial. Typically we do not have access to that distribution and hence we resort to alternative ways of reconstructing the vicinity of an image. While LIME assumes that the neighbourhood of an image can be approximated by sampling a subset of patched versions of that image, we propose this is not the best approximation. In this Section, we first illustrate how to improve upon this restriction by sampling from the real distribution in the context of a toy problem in 2D. Then, we proceed to show that similar results can be achieved in the visual domain. However, as we do not know the distribution of all possible real images, rather than sampling from it, we need to approximate it. Furthermore, we explore the use of perceptual metrics as an alternative proxy to enforce a similar behaviour. In all cases, a ridge regressor is used as our surrogate model.

4.1 Synthetic example

First, we test how altering the sampling can affect the resulting surrogate explainer in a simple 2D case. We use a non-parametric uniformisation transform in order to achieve a gradual scale between sampling independent of the distribution using a bounded uniform distribution around our query point, and according to the original data distribution. We explore the impact this has on the resulting surrogate in a dimensionality which we can visualise before presenting experiments using real images. We use a two moons dataset, with additive Gaussian noise with a standard deviation of 0.35 to ensure the classes are linearly inseparable. A random forest is trained on samples taken from this distribution.

Approximating the data distribution In order to achieve a scale between sampling according to the distribution or independently of it, we use a quantile transformation. An estimate of the cumulative distribution function is used to map the values to a uniform distribution. The number of quantiles, or the number of points the CDF is estimated using, dictates the degree of uniformisation. A high number of quantiles leads to a more accurate estimation of the CDF and a more uniform transformation. To sample our data for the surrogate explainer, we initially sample uniformly within certain bounds around the query point in order to enforce locality. We then progressively use the inverse transform of the quantile uniformisation, with an increasing number of quantiles. The result is data that is transformed to be more like the original distribution the higher the number of quantiles used. Surrogate explainers are then obtained using this sampled data, and we observe how different these surrogates are compared to a surrogate trained on data from local samples from the true distribution.

Evaluation The Wasserstein distance is measured between the sampled data and samples from the true distribution within the same bounds to measure similarity between the sampled data and the distribution. The ℓ_2 distance between parameters of the ridge regression are used to track the distance between the obtained explainer, and the surrogate explainer trained on the true distribution.

Experiments We first train the quantile uniformisation transforms, with quantiles in the range $[2, 100]$. 50 query points x are chosen from a test set of the two moons distribution,

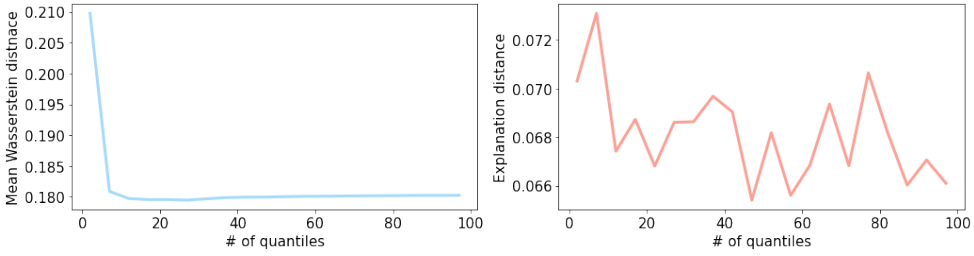


Figure 4: Mean Wasserstein distance and explanation distance as a function of the number of quantiles used to transform the synthetic data.

and we sample uniformly around them in both feature directions in the range $[x-\sigma, x+\sigma]$, with $\sigma = 0.2$ to ensure that the samples and the resulting explanation are local. The inverse uniformisation is then applied (going from a uniform distribution to an approximation of the two moons distribution) and a surrogate is trained on this data.

Results Fig. 5 is an example of the generated surrogates for a point close to the global model decision boundary and towards the edge of the distribution. Fig. 4 shows the average across 50 query points. As we increase the number of quantiles used in the transformation, the Wasserstein distance decreases. This means our sampled data is more similar to the original distribution. The distance between the resulting surrogate and one trained on the true distribution also decreases, meaning that when we sample data increasingly more similar to the the true distribution, the resulting surrogate explainer is also more similar. This may seem somewhat trivial, but it is important to note that this concept can be applied to more complicated distributions, where we cannot sample data easily and then transform according to the distribution.

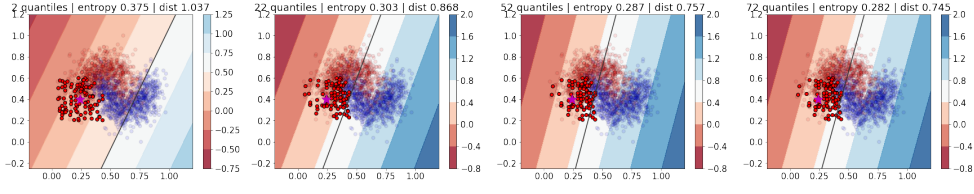


Figure 5: Surrogates trained on the two moons dataset, where the sampled data is transformed to be increasingly similar to the original distribution as we move to the right. Left image resembles uniform sampled data.

4.2 Natural images

Dataset TID2008 is a dataset composed of 25 natural images depicting different objects, landscapes and people, as well as some artificially-generated scenes [18]. It was originally devised to evaluate visual quality assessment metrics in images affected by 17 types of distortion (like adding salt&pepper noise or having compression errors) at 4 different degrees of intensity. Due to the computation expenses associated with generating an explanation, we restrict our experiments to clear samples and their counterparts containing either additive

Gaussian noise, Gaussian blurring or a change in contrast. Hence, we create an explanation for each of the 100 pictures available, for each type of distortion considered. Images are downsampled to 224 x 224 pixels in all our experiments.

Approximating the data distribution Superpixels remain the interpretable data representation in our experiments, yet we extend on the meaning of the binary mask of the original LIME. In order to mimic the data the black-box model is trained from, instead of setting pixels within a superpixel to their mean value, our binary vectors represent whether pixels are altered according to one of the distortions considered in our dataset (Figure 3). We either apply additive Gaussian noise with increasing mean intensity values 0.01, 0.05 and 0.1, Gaussian blurring with kernel sizes 3×3, 5×5 and 11×11 or a change in the contrast of the image. It should be noted that the more intense the distortion is, the information within the superpixel becomes more masked, but the image lies further towards the edge of the distribution of natural images. In this case, a change in contrast of 0.5 halves the original contrast level of an image, whereas a value of 1 leaves it intact. Considering the nature of the images to explain and modifying the sampling accordingly, we are leveraging the resources available to make the training data of the surrogate more realistic, as well as more similar to that of the black-box training data.

Black-box model Transformer architectures, although initially introduced to tackle natural language processing tasks [3, 29], are also becoming a popular choice in the computer vision domain [21]. Due to their adoption by the community, we choose a Vision Transformer (ViT) model trained for image classification as our black-box predictive system [4]. The specific implementation we use is a version pretrained on the Imagenet-21K dataset [25].

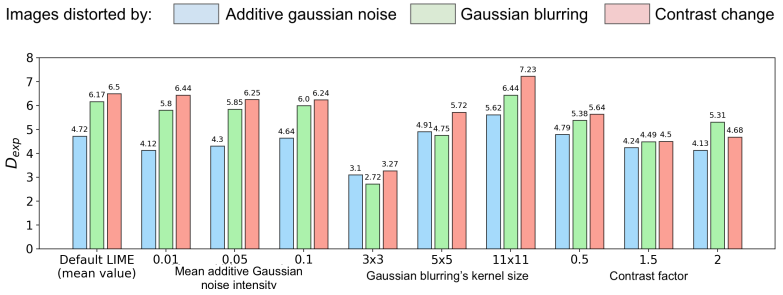
Distance between explanations The evaluation of the quality and reliability of computational explanations of black-box models is a complex problem [20, 27]. To judge the similarity between explanations, we follow the procedure introduced in [9], where the authors define the empirical distance between two explanations $\mathbf{E_k}$ and $\tilde{\mathbf{E}}_\mathbf{k}$ as

$$D_{exp} = \frac{1}{K} \sum_{k=0}^{K} |\mathbf{E}_k - \tilde{\mathbf{E}}_\mathbf{k}|_F^2 \tag{4}$$

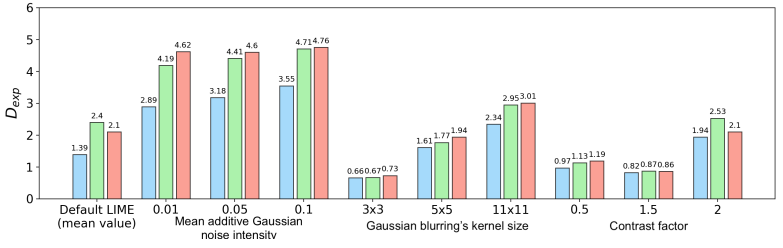
where the sum is performed over K explained classes. In our experiments we generate surrogates only for the most likely class predicted by the classification model, thus $K = 1$. $\mathbf{E_k}$ (and analogously, $\tilde{\mathbf{E}}_\mathbf{k}$) is a matrix in which each $E_k(i, j)$ denotes the importance value in the explanation of the superpixel the pixel in position (i, j) belongs to for class k.

Experiments Surrogates for pairs of reference and distorted images are trained using the same combination of sampling procedure and distance metric, to ensure coherence in the explanations generated. These are then projected as pixel relevance maps and the distance between them is computed using the definition above. We repeat this procedure for all reference images and average over each type of distortion considered.

Results First of all, it can be seen that by using a perceptual distance to measure the proximity of the local neighbours, we can partially alleviate the lack of direct access to the training data distribution as seen by the black-box model. It can be noticed that by weighing



(a) Using a cosine distance metric.



(b) Using MS-SSIM as distance metric.

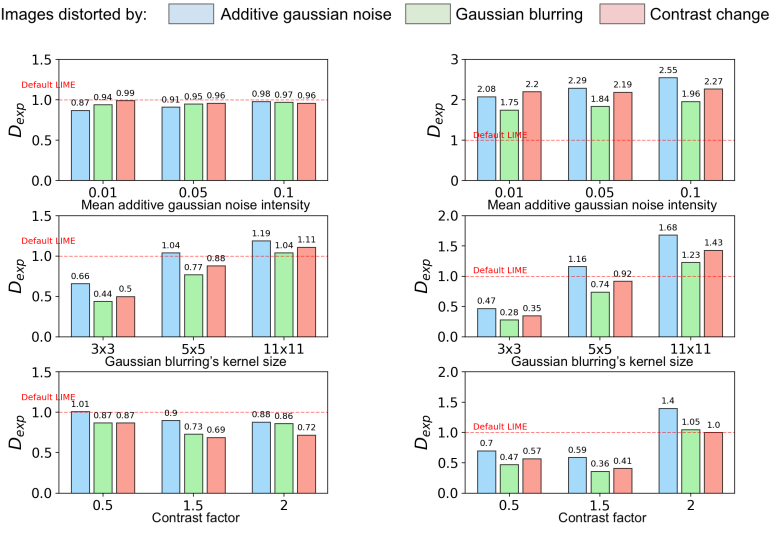
Figure 6: Average absolute distance computed for each pair of clear-distorted images depending on the sampling method, the distortion type of the image and the distance metric used. The *x*-axis indicates the transform applied to superpixels in the sampling process.

samples according to MS-SSIM (Fig. 6b), the explanations generated for an image and a distorted version of it are always closer on average than using a cosine distance (Fig. (6a). The explanations are more robust in the pixel space, in spite of the type of distortion or particular space we sample from. The same distance values, but normalised to the value of the distances computed by a default LIME configuration, are shown in Fig. 7. It is worth pointing out that if we use a cosine distance, aligning the sampling method to the distortion type of an image results in smaller or equivalent distances than those using the classical LIME approach or other sampling procedures. Although this is true in the case of the cosine distance, the same does not seem to hold if we use MS-SSIM.

5 Conclusions and Future Work

Post-hoc surrogate-based explanation methods like LIME, although widely adopted, rely on sampling from a neighbourhood around a query in order to approximate the local decision boundary learned by a non-interpretable model. This does not take into account the data distribution with which the black-box model was trained, even though we would ideally sample the neighbourhood to train surrogates from that distribution.

In this paper, we showed that for distributions simple enough to be estimated with traditional density estimation techniques, like our 2D example, sampling from the estimated distribution provides a reasonable vicinity to train a surrogate. In order to test the same idea on the visual domain, due to the intractability to directly access the true distribution from which the training dataset was sampled (or the distribution of all real images), we proposed two complementary approaches to approximate that distribution locally.



(a) Cosine distance (b) MS-SSIM

Figure 7: Average distance between explanations normalised to the mean distance computed using the default LIME configuration. It is important to align some sampling methods to the nature of the images explained in the absence of access to the true distribution.

On the one hand, we demonstrated that perceptual metrics, as they implicitly convey information about the distribution of real images, can be successfully used as a proxy to the distribution of real images, improving the consistency between explanations for similar images. On the other hand, rather than sampling from patched versions of a query, as is customary, we sampled from distributions of images undergoing some degree of distortion that resemble more closely the set of real images. We found that in some cases we can further improve the robustness of explanations when combining an aligned sampling method with both euclidean and perceptual distances.

Future research will be devoted to create neighbourhoods not only from interpretable domains based on the visual features of the query, but also on its semantics. To that end, generative models like DALL-E2 [22] or Imagen [26], both could be embedded within our pipeline for realistic semantically-driven in-painting of images to better approximate the manifold of real images.

6 Acknowledgements

The work leading to these results was supported by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22, funded by MCIN/AEI/10.13039/501100011033), and AMIC-PoC (PDC2021-120846-C42, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR"). Ricardo Kleinlein's research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225). This work was also funded by the UKRI Turing AI Fellowship EP/V024817/1.

References

[1] Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.

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

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[4] 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. *ArXiv*, abs/2010.11929, 2021.

[5] Hui Fen, Kuangyan Song, Madeilene Udell, Yiming Sun, Yujia Zhang, et al. Why should you trust my interpretation? understanding uncertainty in lime predictions. *ArXiv:1904.12991*, 2019.

[6] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *Proceedings of the IEEE International Conference on Computer Vision*, October:3449–3457, 2017.

[7] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[9] Alexander Hepburn and Raul Santos-Rodriguez. Explainers in the wild: Making surrogate explainers robust to distortions through perception. *International Conference on Image Processing (ICIP)*, September:3717–3721, 2 2021.

[10] Alexander Hepburn, Valero Laparra, Raul Santos-Rodriguez, Johannes Ballé, and Jesus Malo. On the relation between statistical learning and perceptual distances. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P. Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 16:1–6, 2016.

[12] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[13] Scott Lundberg and Su-in Lee. A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 32(2):1208–1217, 2019.

[14] Jesús Malo and Juan Gutiérrez. V1 non-linear properties emerge from local-to-global non-linear ica. *Network: Computation in Neural Systems*, 17(1):85–102, 2006.

[15] Jesús Malo and Valero Laparra. Psychophysically tuned divisive normalization approximately factorizes the pdf of natural images. *Neural Computation*, 22:3179–3206, 2010.

[16] Marina Martinez-Garcia, Praveen Cyriac, Thomas Batard, Marcelo Bertalmío, and Jesús Malo. Derivatives and inverse of cascaded linear+ nonlinear neural models. *PloS one*, 13(10):e0201326, 2018.

[17] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[18] Nikolai Ponomarenko, Vassili Lukin, Ander Zelensky, Kiril. Egiazarian, Mona Carli, and F. Battisti. Tid2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009.

[19] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[20] Rafael Poyiadzi, X. Renard, Thibault Laugel, Raúl Santos-Rodríguez, and Marcin Detyniecki. On the overlooked issue of defining explanation objectives for local-surrogate explainers. *ArXiv*, abs/2106.05810, 2021.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *The 38th International Conference on Machine Learning (ICML)*, 2021.

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (ICKDD)*, 2016.

[24] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.

[27] Kacper Sokol and Peter Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *Kunstliche Intelligenz (KI)*, 34: 235–250, 2020.

[28] Kacper Sokol, Alexander Hepburn, Raul Santos-Rodriguez, and Peter Flach. blimey: surrogate prediction explanations beyond lime. *arXiv:1910.13016*, 2019.

[29] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[30] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:841, 2017.

[31] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multi-scale structural similarity for image quality assessment. *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, 2:1398–1402, 2003.