Sampling Based On Natural Image Statistics Improves Local Surrogate Explainers

Ricardo Kleinlein[1], Alexander Hepburn[2], Raúl Santos-Rodríguez[2], Fernando Fernández-Martínez[1]

[1]Information Processing & Telecommunications Center, ETSIT, Universidad Politécnica de Madrid, [2]Department of Engineering Mathematics, University of Bristol

Contact: ricardo.kleinlein.at.upm.es

Abstract

Surrogate explainers are a popular **post-hoc interpretability method** to understand how a black-box model arrives at a particular prediction. Interpretable domains have traditionally been derived exclusively from the intrinsic features of the query image, ignoring the data distribution used to train the black-box model. This leads to **surrogates** trained on **images that lie within low probability regions of the manifold of real images**.

We propose to approximate the **original training data distribution, even when this distribution is not accessible**: (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.

Post-hoc local surrogate explainers

Approximations to the boundary decision of a black-box system around a query point x (usually linear) through a sampled neighbourhood $\boldsymbol{\pi}_x$:

$$\exp_x = \underset{g \in \mathcal{G}}{\arg\min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Interpretable data representation

Sampling in a neighbourhood is done in a human interpretable domain; superpixels and masking regions of the image.

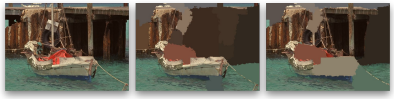
Sampling a neighbourhood

These binary vectors denote whether pixels within a superpixel undergo mean color occlusion. To measure the distance between the generated samples and the query x:

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

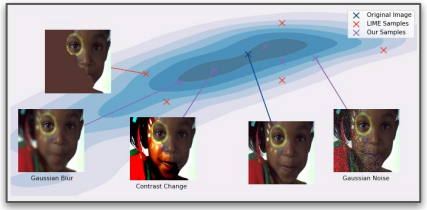
Natural image statistics & LIME [1, 2]

To build an interpretable domain and train the local surrogate, a neighbourhood is sampled around the query image. **Current sampling techniques** used in the most popular surrogate explainer method (LIME) **create samples that are far from what we can consider real images**.



(1) Sampling from distortions

Ideally, we would sample from the actual distribution of real-world images. Instead, we approximate it by **transforming pixels according to distortions** the human eye is particularly sensitive to.



(2) Perceptual metrics - MS-SSIM [3]

Distances that attempt to reflect the **similarity between images as it is perceived** by the human visual system.

Results on natural images

Dataset: TID2008 (subset)

25 natural images subject to 3 distortions (Gaussian noise/Gaussian blurring/Contrast) at 4 levels of intensity.

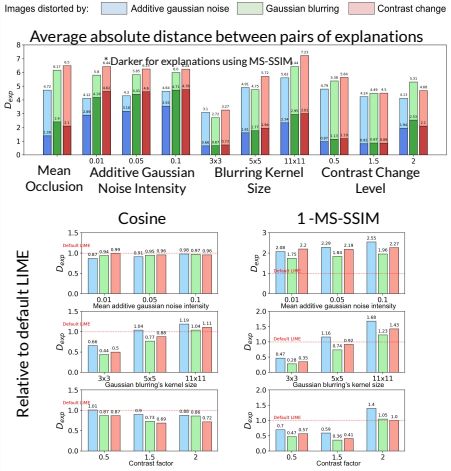
Black-box model

Vision Transformer (ViT) model trained for image classification, pretrained on the Imagenet-21K dataset.

Distance between explanations

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

E: Explanations, K: Number of classes predicted (K=1)



Conclusions & Future work

- For **distributions simple enough to be estimated** with traditional density estimation techniques, **sampling from the estimated distribution** provides a reasonable vicinity to train a surrogate.

- Since the **distribution of natural images is not known or intractable**, **perceptual metrics** seem to convey information about true distribution.

- When **perceptual metrics are aligned with perceptually-meaningful sampling** methods, it can boost the robustness of explanations.

Still, neighbourhoods should be created not only from the visual features of the query, but **also on its semantics**. Generative vision models might be used to sample realistic training samples for the surrogate in the vicinity of the query point..

Main references

[1]: 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.

[2]: 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.

[3]: 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.

Acknowledgments

The work leading to these results was supported by the Spanish Ministry of Science and Innovation: GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22), AMIC-PoC (PDC2021-120846-C42), the Spanish Ministry of Education (FPI grant PRE2018-083225) and also by the UKRI Turing AI Fellowship EP/V024817/1.