SLICE: Stabilized LIME for Consistent Explanations for Image Classification

Supplementary Material

S1. Overview of LIME

LIME is a popular method for interpreting the predictions of complex machine learning models introduced in 2016 [13]. The main idea behind LIME is to approximate a complex model locally with a simpler, transparent model (like linear regression or decision trees), thus providing an explanation for individual predictions. Mathematically, LIME aims to solve the following optimization problem:

$$\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{5}$$

Where, $f(x)$ is the prediction of the complex model, for instance, x, $g(x')$ is the prediction of the surrogate model for a representation x' of instance x, $\pi_x(z)$ is a proximity measure between instance x and z and $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithfully g approximates f in the vicinity of x, weighted by the proximity measure $\pi_x(z)$ and $\Omega(g)$ is a measure of the complexity of the surrogate model.

The optimization aims to find a surrogate model g that both approximate the complex model f well in the vicinity of the instance x and is transparent in nature. The key to LIME is the choice of the representation x' and the measure of locality $\pi_x(z)$. In the original paper, the authors propose using a binary vector x' that indicates the presence or absence of interpretable components (like words in a text or superpixels in an image), and a measure of locality that gives higher weight to instances that are closer to x. The weight function is an exponential kernel, i.e. $\pi_x(z) = \exp(-D(x, z)^2/\sigma^2)$, where $D(x, z)$ is the cosine distance between x and z, and σ is a kernel width parameter.

In the context of image data, this involves transforming the image problem into a tabular format as shown in Fig. S1. The process has the following main steps viz. 1) Dividing the image into superpixels using segmentation, 2) Generating random perturbation vectors with length equal to a number of superpixels, 3) Perturbing the superpixels and noting the predictions 4) Building a surrogate model with perturbation vectors as X and predictions from step 3 as Y, and 5) extracting explanations from the surrogate model. This transformation of the image problem into a tabular format is a key part of how LIME provides explanations for image classification models.

S2. Details of Statistical Tests

We performed several tests to ascertain the statistical significance of our results. We performed the Wilcoxon rank test to confirm that the CCM scores of SLICE were higher than those of LIME and BayLIME. The low p-values and high Test Statistics in Tab. S1 provide robust statistical evidence.

Fig. S2a and Fig. S2b show the distribution of the difference of the AOPC score using the deletion process. Most differences are above 0, indicating that the AOPC score using the deletion process for SLICE explanations is higher than that of LIME and BayLIME. The same is observed in Fig. S3a and Fig. S3b for the AOPC score using the insertion process.

Table S1. Wilcoxon rank test results for comparison of LIME, BayLIME, and SLICE, SLICE_blur and SLICE_FE. Here x,y in the test column indicates the test details with x and y. Where x and y are one of S, Sb, Sf, L and B and denotes SLICE, SLICE_blur, SLICE_FE, LIME and BayLIME respectively. The null hypothesis H_0 was "The median of the differences ($CCM(x) - CCM(y)$) is equal to zero," and the alternative hypothesis was H_a was "The median of the differences ($CCM(x) - CCM(y)$) is greater than zero". D:M denotes Dataset:Model where O refers to Oxford-IIT Pets and P refers to PASCAL VOC datasets. R denotes Resnet50 and I denotes Inception V3 models. W denotes the Test Statistic, M_Δ denotes the median of differences and Neg. Count denotes the number of negatives out of 50 images

Test	D:M	W	p-value	M_Δ	Neg. Count
S, L	O:R	1271	4.8e-10	0.65	1
S, L	O:I	1275	3.8e-10	0.76	0
S, L	P:R	1275	3.8e-10	0.60	0
S, L	P:I	1275	3.8e-10	0.60	0
S, B	O:I	1275	3.8e-10	0.67	0
S, B	O:R	1230	5.3e-09	0.50	1
S, B	P:I	1275	3.8e-10	0.67	0
S, B	P:R	1275	3.8e-10	0.48	0
S, Sb	O:I	1266	6.5e-10	0.20	2
S, Sb	O:R	1092	9.2e-07	0.02	6
S, Sb	P:I	1247	2.0e-09	0.22	3
S, Sb	P:R	990	3.3e-4	0.01	10
S, Sf	O:I	1275	3.8e-10	0.53	0
S, Sf	O:R	1261	8.8e-10	0.51	1
S, Sf	P:I	1275	3.8e-10	0.56	0
S, Sf	P:R	1275	3.8e-10	0.51	0

Fig. S4a and Fig. S4b show the distribution of the difference of the AUC using the deletion process. Most of the differences are above 0, indicating that the AUC scores using the deletion process for LIME and BayLIME explanations are higher than those of SLICE. Unlike the other statistical tests, the differences are calculated by subtracting the AUC scores of SLICE from that of LIME or BayLIME. This is because, for the deletion process, a lower AUC score indicates higher fidelity.

In Fig. S5a and Fig. S5b shows the distribution of the difference of the AUCs using the insertion process. Most of the differences are above 0, indicating that the AUC scores using the insertion process for SLICE explanations are higher than those of LIME and BayLIME.

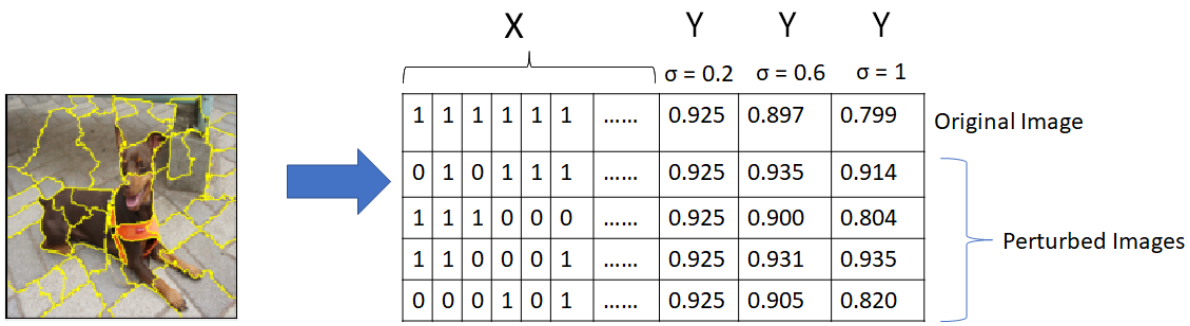
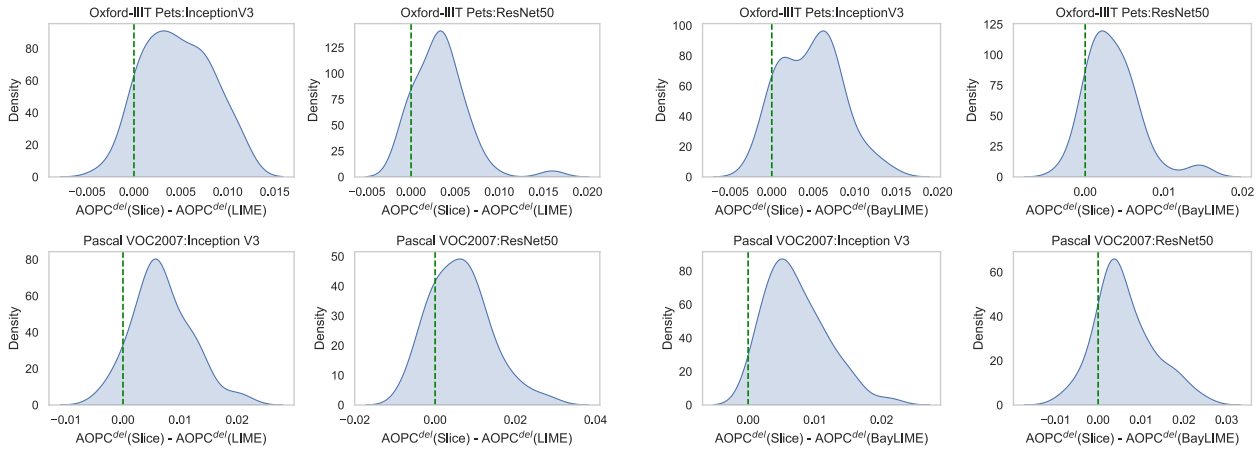


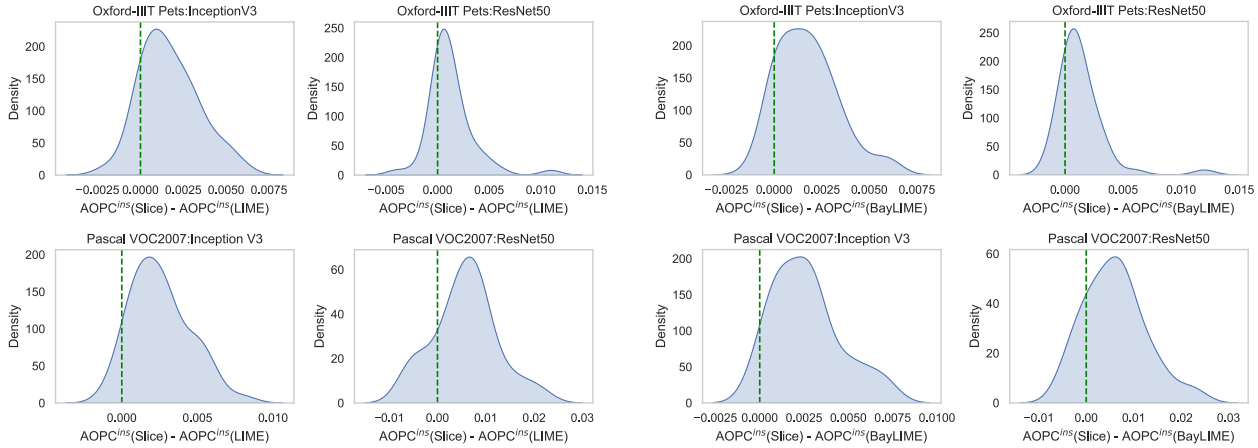
Figure S1. Adaptation of LIME transformation for selecting sigma of Gaussian blur



(a) SLICE - LIME

(b) SLICE - BayLIME

Figure S2. Plot of differences of AOPC scores for deletion process



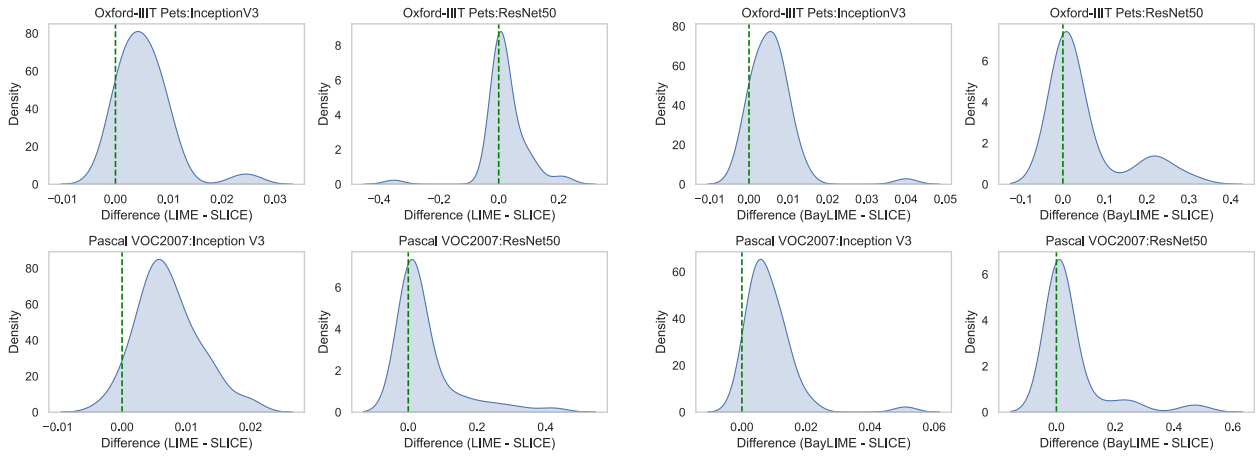
(a) SLICE - LIME

(b) SLICE -BayLIME

Figure S3. Plot of differences of AOPC scores for insertion process

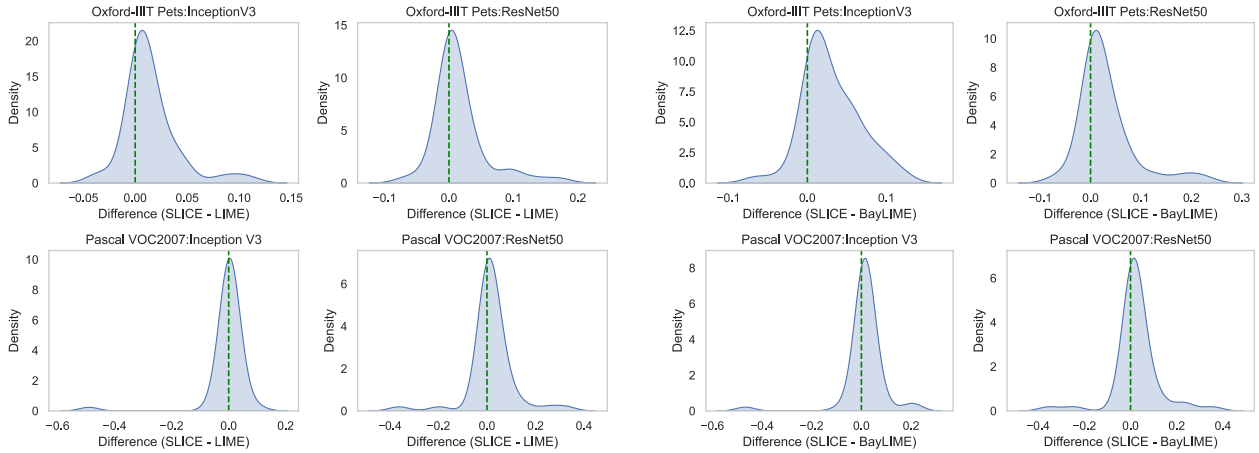
S3. Visualization of Explanations

Fig. S6 shows the explanations from LIME, BayLIME and SLICE. The top row shows LIME explanations from 4 random runs while the middle row shows the same for BayLIME. The bottom row shows the explanations from SLICE. The superpixels highighted in blue are those superpixels which the respective method deemed as positively influencing the output probability. Whereas, the su-



(a) LIME - SLICE (b) BayLIME - SLICE

Figure S4. Plot of differences of AUCs for traditional deletion game



(a) SLICE - LIME (b) SLICE - BayLIME

Figure S5. Plot of differences of AUCs for traditional insertion game

perpixels highighted in red are those superpixels which the respective method deemed as negatively influencing the output probability. LIME explanations show inconsistency in terms of Sign Entropy and also the superpixel importance ranks. Some of the superpixels explained as contributing positively to the output probability for the class Egpytian Cat are also shown as making negative contribution in another random run. Further, the importance ranks of the super pixels also vary. For BayLIME there is no consistency problem arising due to Sign Entropy as there is no overlap of positive and negative superpixels. However, the superpixel ranks vary among themselves within the positive and negative sets. SLICE, on the other hand, shows consistency both in Sign Entropy and importance ranks of superpixels. The positive superpixels in LIME and BayLime explanations scattered within the region of the cat and also at times points to the back ground. However, the positive superpixels in SLICE explanations are strictly within the cat. The fidelity results from Sec. 5.3 and Sec. S2 prove that SLICE explanations have higher fidelity. Hence, it can be concluded that some explanations from LIME and BayLIME may deviate from the actual regions the model is considering for decision making.

S4. Scott's Rule for Bandwidth Selection

The bandwidth (h) according to Scott's rule is given by:

$$h = 1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}$$

Where:

- n is the sample size.
- $\hat{\sigma}$ is the standard deviation of the dataset, serving as an estimate of the population standard deviation.

(a) Highlighting the top 5 positive features identified by LIME, BayLIME and SLICE for Inception V3 model with a sample image with from Oxford-IIIT Pets dataset

(b) Highlighting the top 5 negative features identified by LIME, BayLIME and SLICE for Inception V3 model with a sample image with from Oxford-IIIT Pets dataset

Figure S6. Representative explanations from four random runs of LIME(top row), BayLIME (middle row) and SLICE(bottom row) for the top predicted class (i.e., Egyptian cat) by Inception V3 model. The blue mask denotes the top five positive superpixels, and the red mask represents the top 5 negative superpixels. The ranks are indicated by white color, and the lower rank signifies higher importance.

S5. Consistency Evaluation on ViT model

Like LIME, our method is independent of model architecture as it needs only the model's input and output. This is different from CAM-based approaches. We experimented on a ViT model trained by Google with patch size of 16×16 and obtained similar results (Refer Fig. S7 below).

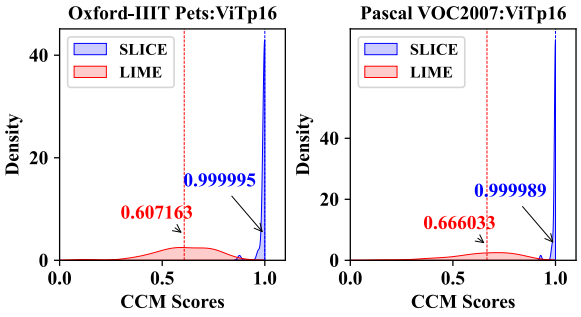


Figure S7. Distribution of CCM Scores for ViTp16 model