Deep Learning Lab-4

immediate

October 17, 2024

Abstract

This Lab focuses to implement two blackbox explainer models, LIME and RISE. These models explain predictions by identifying the pixels or regions that most influence the output, without needing access to the modelâs architecture or weights. They rely only on the input image, the classification score, and the predicted class.

Both methods generate saliency maps by perturbing parts of the image and observing how these changes affect the prediction. The idea is that important features, when altered, should significantly impact the modelâs output, allowing the explainer to highlight the most relevant areas of the image.

1 Explanation Models

Explanation models clarify the decision-making process of complex models, making their predictions more understandable. They provide insights in a way humans can easily interpret, helping to make blackbox models more transparent.

1.1 Blackbox models

Opaque models, often referred to as blackbox models, make predictions in a way that is not easily interpretable by humans. Their internal logic, rules, and processes are complex and opaque, making them difficult to understand.

1.1.1 LIME

LIME is a feature attribution method that determines the importance of each feature in an input sample for a modelâs prediction. It works by perturbing the input sample and analyzing the corresponding changes in predictions to identify important features.

1.1.2 RISE

RISE (Randomized Input Sampling for Explanation) generates a saliency map to highlight the important features for a modelâs prediction. It creates numerous low-resolution binary masks, upscales them, and evaluates the modelâs prediction for each masked input. Unlike other methods, RISE assumes unmasked features are crucial for the prediction if the model's output remains strong after masking.

2 Dataset

The dataset consists of 102 images, equally split between two classes: African Elephants and Black Bears, captured in various environments.

Question 1

As you can see in the previous cell, many parameters must be set manually according to the model and data. Try to identify the right combination of parameters to explain the prediction of the given image (in this case, an African elephant). Experiment with different parameter values and demonstrate which combination provides the best visual explanation.

Parameter	Default Values	Test 0	Test 1	Test 2	Test 3	Test 4	Test 5
top_labels	1	1	1	1	1	1	1
hide_color	[0, 0, 0]	[0, 0, 0]	None	None	[0, 0, 0]	[0, 0, 0]	None
num_lime_features	100000	500	2000	100	100	1000	500
num_samples	5000	2000	5000	3000	3000	7000	3000
num_superpixels	15	8	15	10	10	10	25
positive_only	True	True	True	True	True	True	True
negative_only	False	False	False	False	False	False	False
hide_rest	True	True	True	True	True	True	True

Table 1: Parameter configurations for LIME experiments with default values.

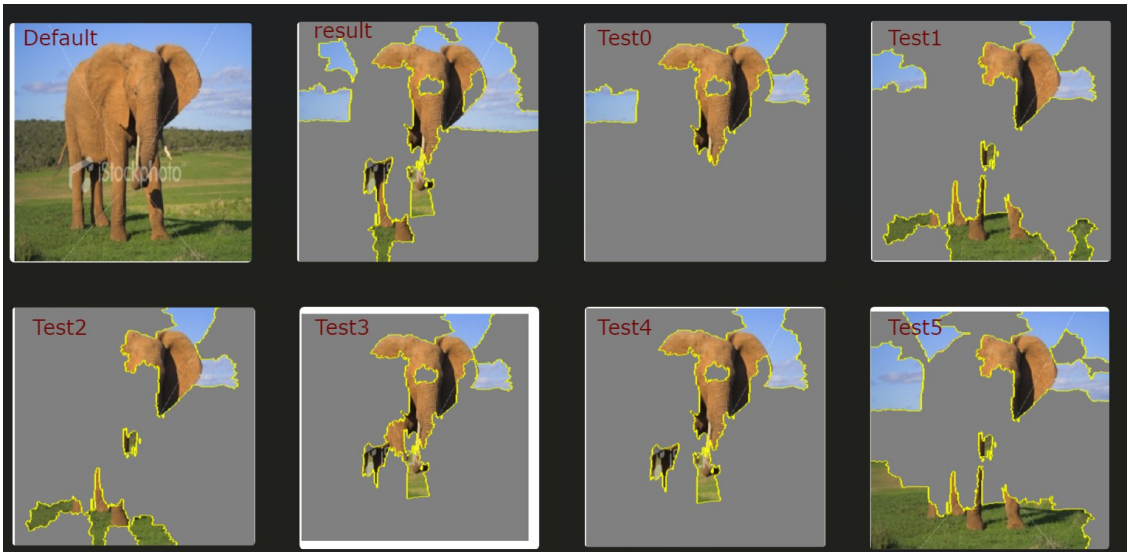

Fig. 1: Lime Result with different configuration.

After experimenting with various parameter combinations, **Test 3** provides the best explanation for the African elephant image. It highlights the relevant parts of the elephant while minimizing background noise. Using 100 LIME features, 3000 perturbation samples, and 10 superpixels, Test 3 achieves the right balance between detail and simplicity, resulting in a clear and focused saliency map.

The top-3 predicted classes for the image are **African Elephant** (49.89% confidence), **Tusker** (29.10% confidence), and **Indian Elephant** (11.87% confidence). This aligns well with the saliency map in Test 3, as it focuses on key elements of the image, such as the elephantâs body and head, which are crucial to the model's prediction.

In contrast, **Test 0** and **Test 1** are too coarse, missing finer details, while **Test 2** and **Test 4** introduce some noise. **Test 5** is overly segmented, making the explanation fragmented. Interestingly, the highlighted regions remain largely consistent across all tests, suggesting the model is robust and stable regardless of LIME settings. The core parts of the image consistently influence the modelâs prediction, with only slight differences indicating a higher influence in certain areas. This further shows that the model is focusing on key features and ignoring less important ones.is the optimal configuration, providing the most comprehensive and accurate explanation without unnecessary noise or over-segmentation.

Question 2

Now, consider another image of an African elephant (see `./data/African_elephant/`). Is your chosen parameter setting still appropriate? How does the output look different for this new image?

Fig. 2: Lime Result with different configuration.

The **Test 3** parameter settings still provide a more accurate and focused explanation for the new elephant image. The model highlights the relevant features of the elephant, such as its body and legs, while avoiding unnecessary background noise, unlike the default parameters. The top-3 predicted classes are **African Elephant** (69.04% confidence), **Indian Elephant** (11.06% confidence), and **Tusker** (8.07% confidence), aligning well with the saliency map in Test 3.

While there are slight differences in the highlighted areas compared to the first image (such as the elephant's feet and water splashes), **Test 3** continues to provide a superior explanation, indicating that the

chosen settings are stable and effective for different images of the same class.

Question 3

Next, we evaluate the chosen parameter setting on images from another class. You can find black bear images at `./data/black_bear/`. What conclusions can you draw based on the results?

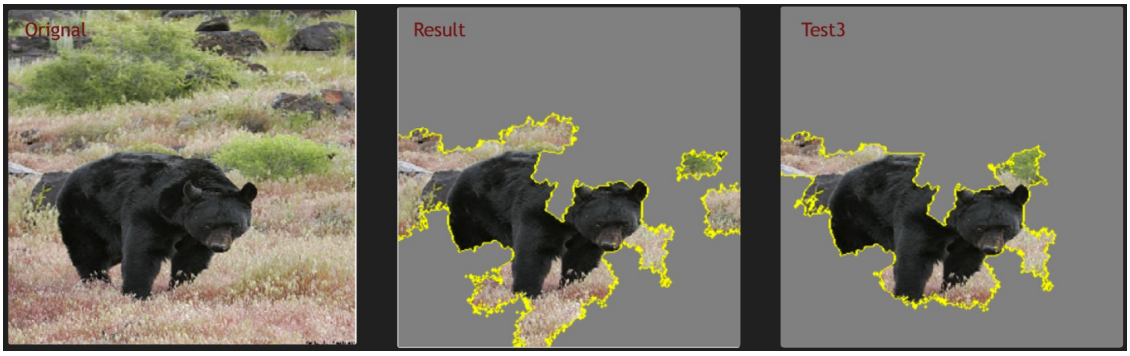

Fig. 3: Lime Result with different configuration.

When applying the **Test 3** parameter settings to an image from a different class (Black Bear), the results remain consistent and focused. The model highlights the relevant features of the black bear, particularly its body and head, while avoiding unnecessary background noise. In comparison, the default parameters also provide good results but introduce more noise, capturing irrelevant areas such as grass and rocks.

The top-3 predicted classes are **American Black Bear** (94.54% confidence), **Brown Bear** (0.31% confidence), and **Sloth Bear** (0.09% confidence), aligning well with the focused saliency map in **Test 3**. Despite the default parameters offering reasonable results, **Test 3** delivers a cleaner and more accurate explanation by minimizing irrelevant background information.

Question 4

In this question, we aim to determine if the parameter setting is more dependent on the data and task or on the model architecture. Specifically, if we change the model, would the same parameter setting still work?

Below is the source code for loading a pre-trained ResNet model. Try to explain its prediction using LIME and identify an optimal parameter setting.

What conclusions can you make from this experiment?

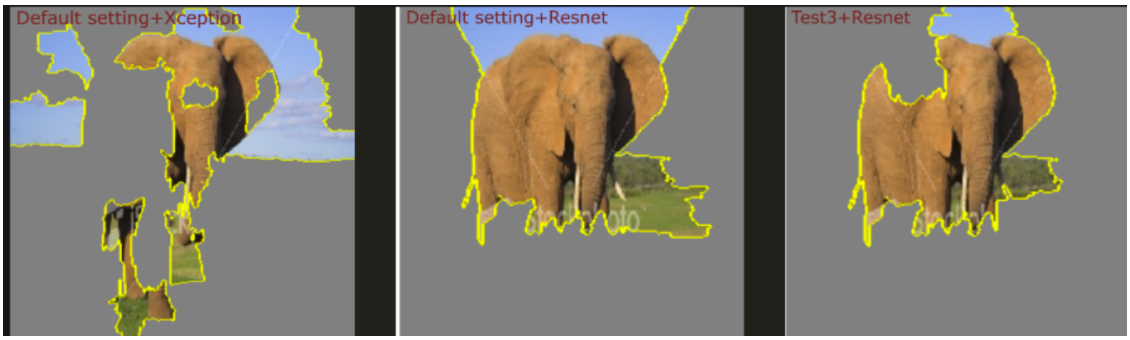

Fig. 4: Lime with RESNET and Xception.

With **Default Setting + ResNet**, the focus on relevant features is significantly improved. The model concentrates on the elephant itself and reduces the noise from the background. The prediction for this model shows a high confidence score for *African_elephant* at **0.7846675**, with much lower scores for *tusker* at **0.20819598**, and *Indian_elephant* at **0.0068448**. This demonstrates that ResNet is better suited for this task with the default parameter setting compared to Xception.

Finally, **Test3 + ResNet** further fine-tunes the feature selection, offering the best result in this comparison. The contours around the elephant are sharper, and the modelâs focus is almost entirely on the key object, further improving the explanation's accuracy.

3 RISE

3.1 Comparison of ResNet and Xception models on an African Elephant image.

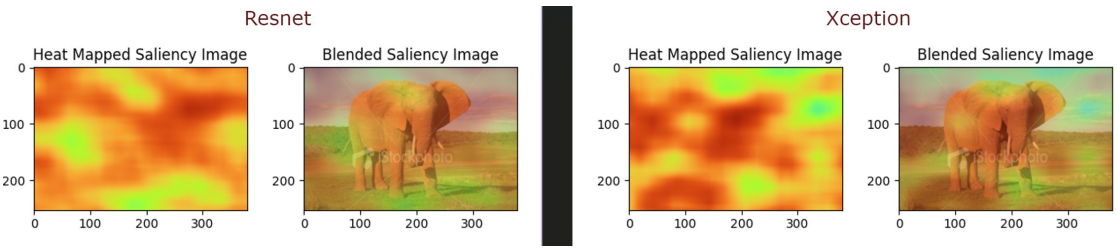

Fig. 5: Rise with RESNET and Xception.

For the particular image, Both models assign saliency across a wide range of the image, but **Xception** appears to focus more on the elephant, particularly its upper body, while **ResNet** spreads its attention more evenly across the entire scene, including the background.

The blended images show that while both models are able to highlight the elephant, **Xception** might offer a cleaner focus on the relevant features, whereas **ResNet** incorporates more of the scene's context.

3.2 Comparison of ResNet and Xception models on an African Elephant image with altered mask values.

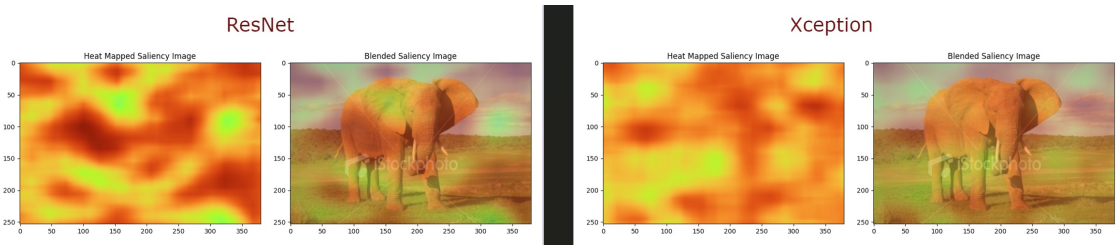

Fig. 6: Rise with RESNET and Xception.

In this case, **ResNet** appears to offer a more refined saliency map, focusing more clearly on the elephant's body, particularly the head and legs, which are highlighted in the heatmap with distinct regions of attention. The contrast between areas of high and low attention is more pronounced.

For **Xception**, the saliency map shows a broader, more evenly distributed attention across the image, with less distinction between regions. The elephantâs body is highlighted, but the attention is spread more diffusely. The difference between the two models' results could be influenced by the **mask number (16)** and **low-resolution mask size (8)**, which affect how precisely the saliency is mapped. ResNetâs map appears sharper, indicating it is more responsive to key features in the image under these parameters.

4 Conclusion

The explanations generated by the two explainer methods, **LIME** and **RISE**, demonstrate that the quality of the results depends on several key factors. These include the stability and generalization capability of the underlying blackbox model, the number of samples, and the feature parameters used in the LIME explainer. For the RISE explainer, the size of the low-resolution mask plays a significant role, with smaller masks typically yielding better explanations though only up to a certain point. Furthermore, the input image itself influences the explanation, meaning that neither explainer method consistently generalizes well across all test images. Therefore, careful tuning of these parameters is essential for obtaining meaningful and reliable explanations.