Explainable AI using LIME

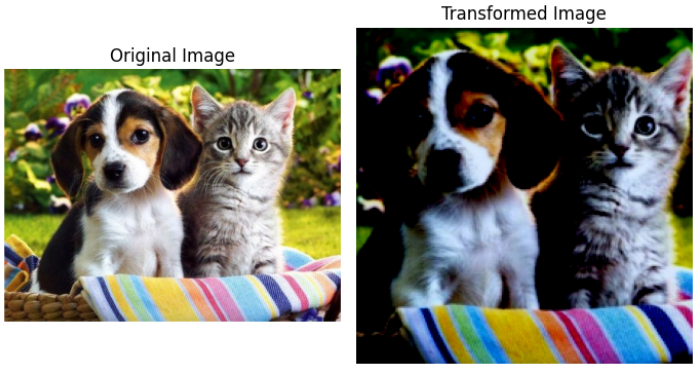
Students Adam Vinestock, Idan Cohen

Introduction

In this report, we showcase the application of LIME (Local Interpretable Model-Agnostic Explanations) on a "black box" image classifying model. Using LIME, we provide local explanations for the top three classes predicted by the model for three selected images.

Choosing Model and Images to be explained

We selected ResNet50 as our image classifier model to be explained. We carefully chose images containing multiple objects to provide clear visual indications for each class. Prior to inputting the images into ResNet50, we applied appropriate transformations to ensure compatibility with the model's requirements.



In order to facilitate the perturbation process on the images, we applied a segmentation technique to divide each image into distinct regions, known as superpixels. Subsequently, we randomly masked 50% of these superpixels. These modified images were then fed into the ResNet50 model, specifically targeting the object class of interest, to obtain new classification scores. To identify the most relevant superpixels contributing to the classification, we employed K-Lasso, a feature selection method. The resulting visualization highlights the superpixels with positive coefficients, providing insights into the discriminative regions impacting the classification outcome.

For image segmentation, we utilized the Scikit Learn SLIC algorithm with a compactness parameter of 10 for superpixel segmentation. We limited the number of segments per image to 20 to balance computational efficiency and runtime considerations. This approach aims to establish a stronger correlation between the linear model's coefficients and the classification scores by ensuring that each object is represented within a small number of segments. Moreover, we employed LASSO feature selection, known for its effectiveness with smaller numbers of coefficients.

Implementing LIME

The interpretable instances created were binary vectors of length 20 representing inclusion/exclusion of superpixels with the corresponding object classification score. Several experiments were conducted to ensure the attainment of compelling visual outcomes. Notably, we observed that the choice of hyperparameters exerted a substantial influence on the final results. The hyperparameters subjected to testing included the following:

- Number of superpixels employed for image segmentation.
- Proportion of the image subjected to masking.
- Count of perturbations executed per image per class.
- L1 regularization coefficient (alpha) utilized in the LASSO algorithm.
- Percentage of superpixels to be exhibited in the final visualization.

Ultimately, our approach involved conducting 300 perturbations per image per class, employing a 50% masking rate, individually adjusting the alpha coefficient for each image, and exclusively showcasing superpixels associated with positive coefficients.

Results

Image1

Image2

Image3

Image1 - Positive Super Pixels for English foxhound

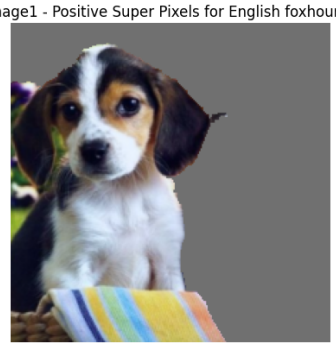


Image2 - Positive Super Pixels for Class analog clock

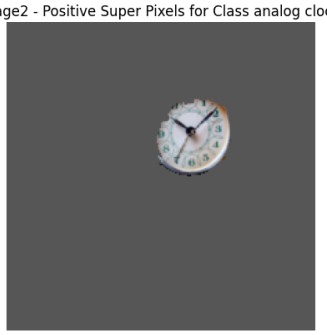


Image3 - Positive Super Pixels for Class knee pad

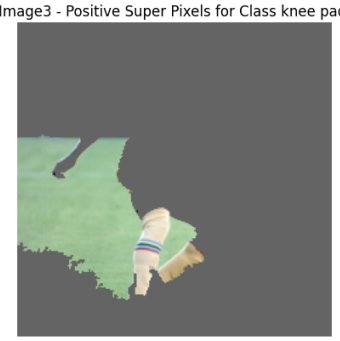


Image1 - Positive Super Pixels for Class beagle

Image2 - Positive Super Pixels for Class barometer

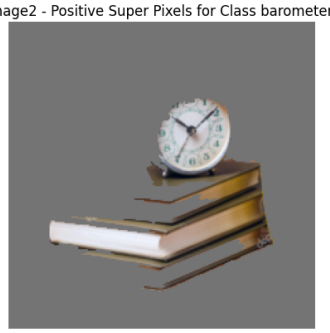


Image3 - Positive Super Pixels for Class rugby ball

Image1 - Positive Super Pixels for Class tabby

Image2 - Positive Super Pixels for Class pencil box

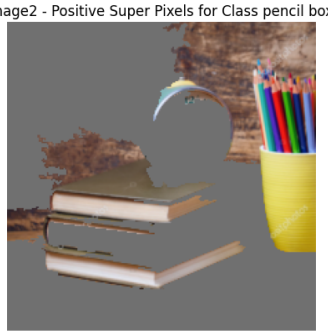


Image3 - Positive Super Pixels for Class soccer ball