Due: July 9th, 2023
CS3946: Advanced Machine Learning

Home Assignment 3: Explainable AI

In this assignment, we explored the LIME method to explain image classification models. We chose a pretrained image classification model and used that to classify sample images. For each of our sample images, we did the following: get the top 3 classes from the model, generated interpretable versions of the images by splitting them to superpixels and represented these interpretable instances as binary vectors (where each entry corresponds to inclusion or exclusion of the superpixels). For each class, we generated a random set of random perturbations of the interpretable instances and then fit a local surrogate model to generate explanations.

LIME

Explainable AI is a rapidly evolving field that aims to provide insights into the decision-making process of machine learning models. As AI models become increasingly sophisticated, it becomes crucial to understand why these models make certain predictions. Explainable AI techniques offer methods to interpret, explain, and gain insights into the inner workings of these models thus enabling users to trust the decisions made by AI systems.

One such technique is Local Interpretable Model-Agnostic Explanations (LIME). LIME is a model-agnostic approach designed to explain the predictions of any machine learning model. LIME focuses on generating explanations at the local level by approximating the model's behavior in the vicinity of a specific instance of interest. LIME's model-agnostic nature enables its application to various machine learning models, enhancing the interpretability and transparency of AI systems across domains.

The process that LIME follows to explain a prediction can be mathematically formulated as solving the following optimization problem:

$\mathcal{L}(f, g, \pi_x)$ - Loss of the approximation of f using g around x

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

In this case: f is the model, g is the approximation of f, π_x is a measure of proximity to the instance being explained x, Ω(g) is a measure of complexity and L is a loss function that measures how well g approximates f. G is the set of possible explanations. The specific definitions of π_x, Ω(g) and L will depend on the problem at hand.
After generating the perturbations and collecting the prediction results, LIME uses these perturbed instances and corresponding predictions to train a local surrogate model. The goal of this surrogate model is to approximate the

behavior of the original complex model as closely as possible around the instance of interest, while being simple enough to interpret. LIME leverages Lasso regression as the local surrogate model due to its feature selection.

Lasso (Least Absolute Shrinkage and Selection Operator) Regression is a form of regularized linear regression. Lasso adds a penalty term to the loss function, which is the absolute value of the magnitude of coefficients. This results in smaller coefficients, which selects a smaller set of input features for use in the model. This property is very useful in the context of LIME because it effectively reduces the number of features that have non zero coefficients.

$$\text{Minimize: } \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In this case, Y_i is the target variable, X_{ij} are the input features, β are the model parameters and λ is the regularization parameters. By utilizing Lasso regression, LIME ensures that the local surrogate model is a good approximation of the original model (in the locality of the instance being explained) and is interpretable (by having a smaller set of important features).

In our assignment, we utilized LIME to explain the predictions of a pretrained image classification model. We selected sample images, obtained the top predicted classes, and generated interpretable versions of the images using superpixels. We represented these interpretable instances as binary vectors and generated random perturbations. By fitting local surrogate models to the perturbed instances, we were able to generate explanations for the image classifications. These explanations provided valuable insights into the model's decision-making process and contributed to our understanding of explainable AI using the LIME method.

Procedure

For each image, we do the following:

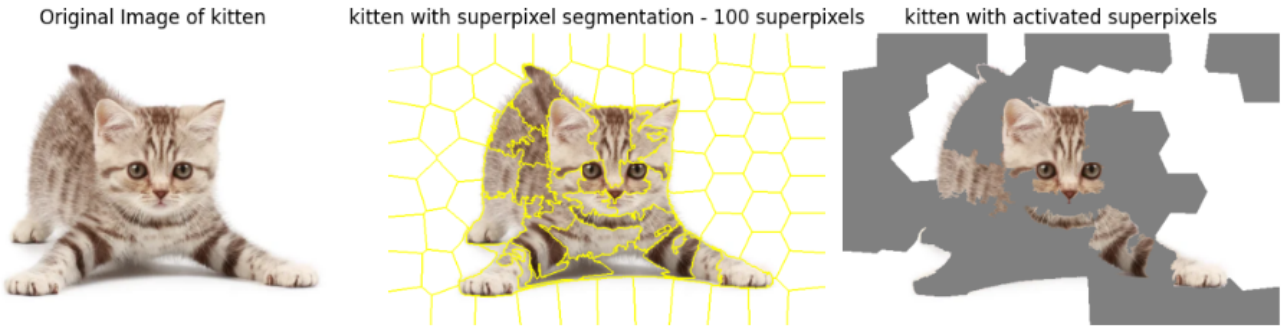
1. Image Classification with ResNet50:
    a. The image classification process begins by utilizing ResNet50, which is a deep CNN known for its ability to accurately classify images across a wide range of categories. This model has been pre trained on a large scale dataset so it already learned to recognize intricate patterns and features within images.
2. Superpixel Segmentation: Display the segmentation by converting the image tensor and applying superpixel segmentation to the image using the SLIC algorithm:
    a. Superpixels are cohesive regions in an image that preserve the image's semantic content in order to gain insight and understand the contribution of different regions.
    b. The SLCI algorithm divides the image into segments based on color and proximity.
3. Binary Mask Generation: Display the binary mask based on the superpixel segmentation of the image, where the mask is created by randomly deactivating a portion of the superpixels:
    a. The masked image is created by assigning gray to the deactivated superpixels in the original image
    b. The binary mask enables the identification of specific areas that impact the models decision making process
4. Perturbation Generation and Interpretability Analysis: Generate a random set of perturbations that create diverse interpretable instances that differ slightly from the original image:
    a. Random perturbations are introduced by deactivating a subset of superpixels
    b. The perturbed image instances provide insights into the effect of selective deactivation on the models classification outcomes
    c. These interpretable instances capture variations in the local decision boundaries
5. Local Surrogate Model Analysis: A local surrogate model is trained to understand the relationship between the interpretable instances and the resulting perturbed scores for each class.
    a. The model leverages the interpretable instances as features and the perturbed scores for each class as the target variable. By analyzing the coefficients of the surrogate model, positive superpixels associated with each class are identified

This procedure allows us to gain a deeper understanding of the image classification process.

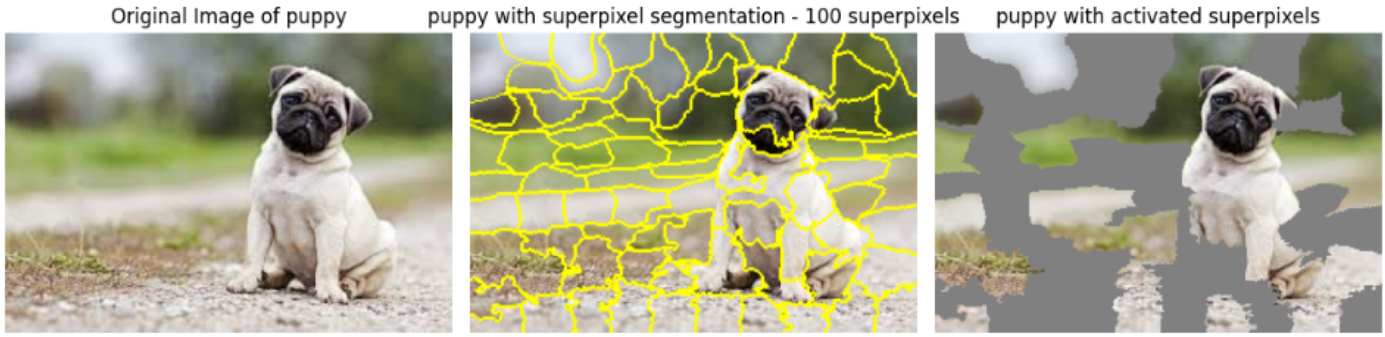
Sample Images

Starting from the original image on the left, the top 3 predicted classes along with their associated scores from the ResNet50 model are returned. The middle image shows the image with superpixel segmentation, which is the result of applying the SLIC algorithm. It partitions the image into nearly uniform superpixels so the original image is overlaid with boundaries that shows the divisions between

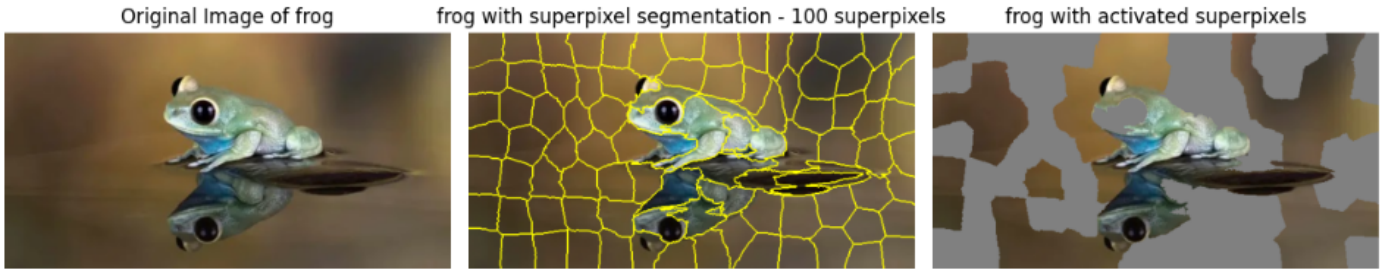
superpixels. The image on the right is the activated superpixels so that only a subset of the superpixels are shown in their original colors (the rest are grayed out). This process highlights which areas of the image (meaning, which specific superpixels) are particularly important for making the prediction.




```
Class scores:
tabby 6.729
Egyptian cat 5.199
tiger cat 4.775
```




```
Class scores:
pug 6.001
Brabancon griffon 4.072
bull mastiff 2.875
```




```
Class scores:
tailed frog 6.318
tree frog 6.177
bullfrog 2.477
```

We generate and display the images that highlight the important superpixels for each predicted class. It does so by generating a local dataset by perturbing the image and calculating the perturbation scores for each class (by deactivating a subset of superpixels in the image then measuring the impact on the classification score). Once the local dataset is generated, we process the image for each predicted class by fitting a linear model (Lasso regression) to identify the important superpixels associated with each class. We display the important superpixels for each predicted class by showing the original image with the remaining important superpixels.

Positive superpixels for tabby

Positive superpixels for Egyptian cat

Positive superpixels for tiger cat

Positive superpixels for pug

Positive superpixels for Brabancon griffon

Positive superpixels for bull mastiff

Positive superpixels for tailed frog

Positive superpixels for tree frog

Positive superpixels for bullfrog

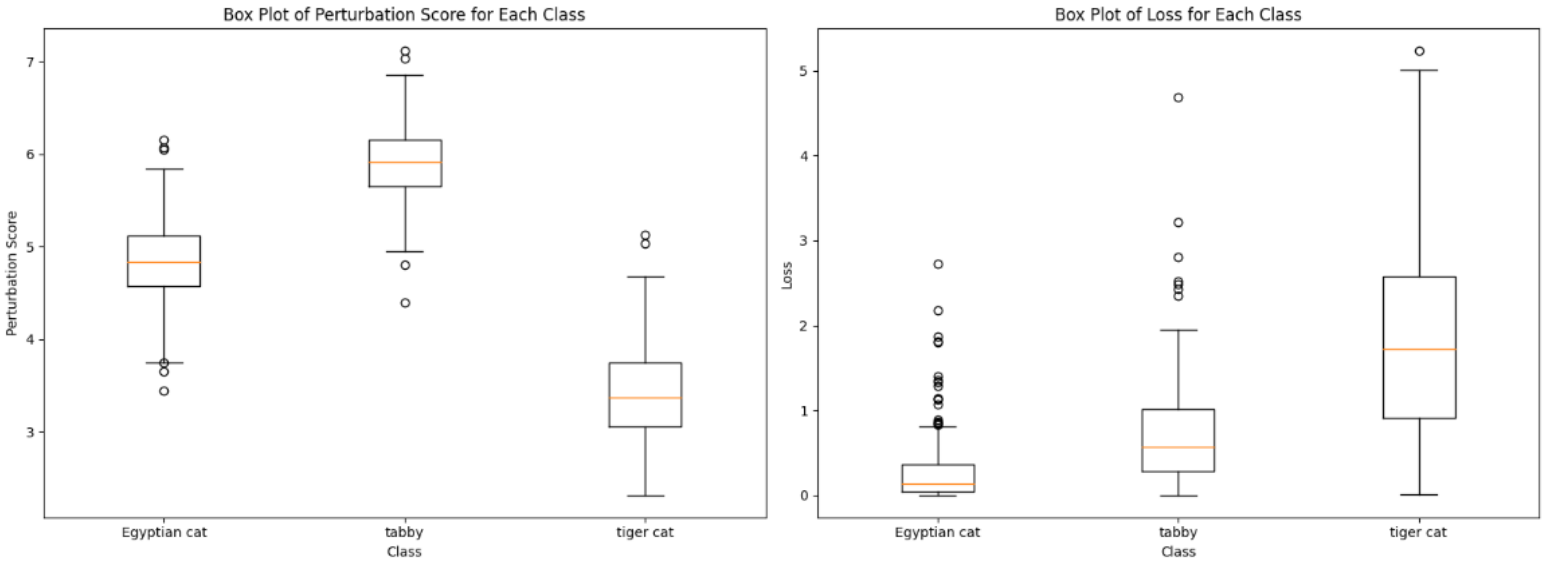
Analysis

The boxplot on the left displays the perturbation scores for each class in the local dataset. The perturbation score is a measure of the sensitivity of the model's prediction to changes in specific regions of the image, and it helps by identifying which superpixels are influential in determining the models decision for a particular class. The boxplot on the right displays the losses for each class. The loss refers

to the value that quantifies the discrepancy between the predicted perturbation score and the actual perturbation score for each class in the local dataset. The loss here provides a measure of how well the local interpretable model (Lasso) fits the data and captures the relationship.

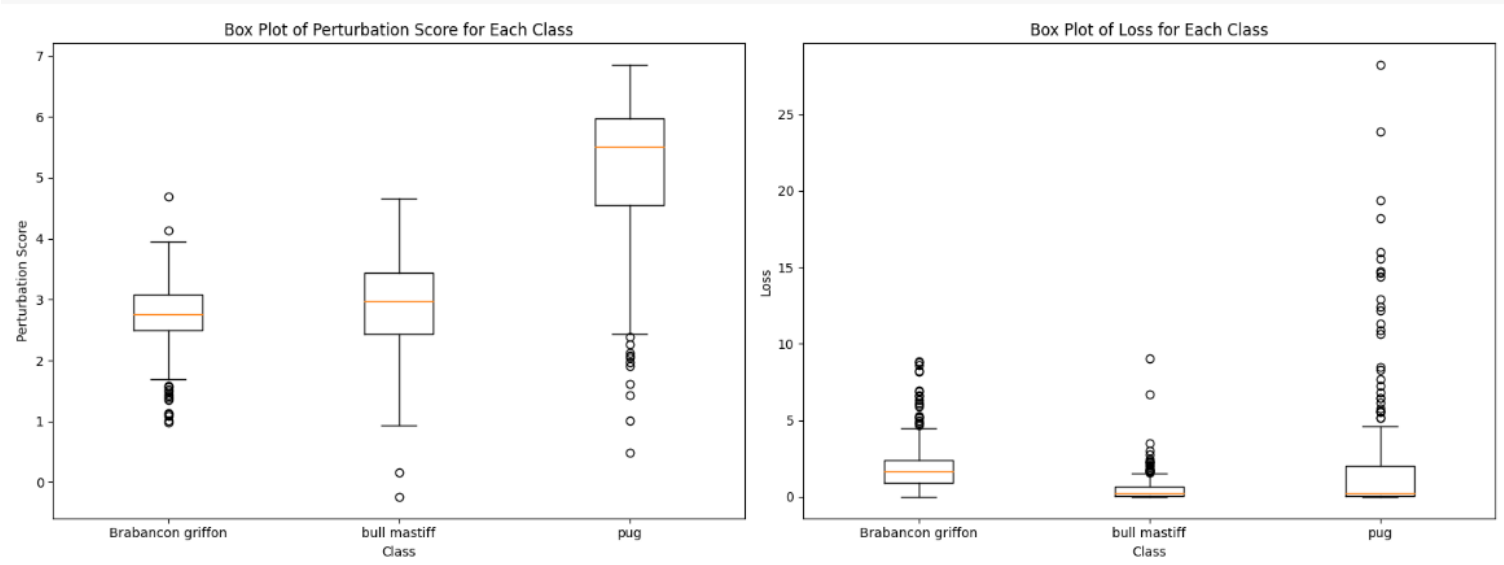
For the cat image, the tabby class (which ResNet predicted with the highest score) has the highest perturbation score which indicates that deactivating the superpixels has a greater impact on the classification score. The tiger cat class (which ResNet predicted with the lowest score) has the lowest perturbation score, which suggests that perturbing the superpixels has less impact on the classification score. The higher perturbation score for the tabby class suggests that specific superpixels play an important role in the models classification for that class, while the lower perturbation score for the tiger cat class indicates the model is less sensitive to changes in those superpixels.

On the other graph, we can see that the tiger cat class has the greater variation which indicates that the relationship between the interpretable instance and the perturbation score is scattered. This indicates a higher level of uncertainty in capturing the importance of superpixels for that class. The greater variation in losses implies that the model's decision making process for this class might involve a combination of various superpixels, rather than relying on a specific set.



For the dog image, the pug class (which ResNet predicted with the highest score) has the highest perturbation score which indicates that deactivating the superpixels has a greater impact on the classification score. The griffon class and the bullmastiff class have comparable perturbation scores, which is surprising given that ResNet predicted griffon with a greater score. This could imply that the model's prediction for the griffon class is not solely dependent on the importance of individual superpixels, but also on other factors such as overall visual appearance or context of the image.

On the other graph, we can see that all three classes have low losses but that pug class has the greatest number of spread outliers while the bullmastiff class has the most concentrated losses. The low losses observed for all three classes indicate that the local interpretable model (Lasso) fits the data relatively well so the relationship between the interpretable instance and the perturbation scores is captured effectively. The tight loss clusters around the median of the bullmastiff class suggest that the relationship between the interpretable instance and the perturbation scores is relatively consistent.



For the frog image, the tree frog and the tailed frog classes have very similar perturbation scores. This makes sense given that their classification score by the ResNet model was close so it aligns with their close classification. The bullfrog has the lowest range of perturbation scores, which makes sense given that it had the lowest classification score. The narrower range indicates that the model is less sensitive to changes in superpixels when distinguishing the bullfrog class, which implies that this classification is influenced by a more limited set of discriminative superpixels.

On the other graph, we can see that the bullfrog and tree frog classes have similar average losses. This similarity indicates that the relationship between the interpretable instance and the perturbation scores is comparable, meaning the model captures a similar level of relationship consistency between superpixels and perturbation scores for both classes.We can also see that the tailed frog has the largest variance which indicates a higher degree of variability in the relationship. The long tail and outliers suggest that there are significant differences among the instances within the data, leading to a more diverse impact of superpixels on the perturbation scores.