Nupur Dhananjay Kulkarni
Léonie Muller

Assignment 4
Task 2 Report

Trustworthy Machine Learning

Analyzing LIME

This report gives an overview of the insights obtained on implementing Task 2 from Assignment 4. The main idea of the report is to analyze the visualizations (for the 10 given ImageNet data points) responsible for the main prediction using a perturbation-based input attribution method called LIME with a ResNet-50 model, with provided weights. LIME is used to explain the individual predictions for the model such that it can be understood which features influence the model prediction the most, for each of the 10 image instances. The visualizations obtained can be accessed here. The implementation can be found here.

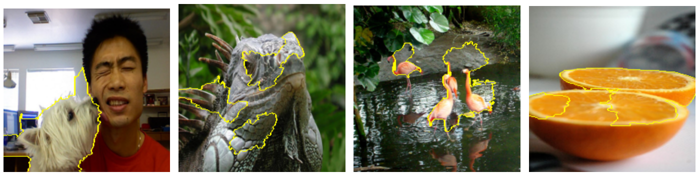

Figure 1: Annotated Images

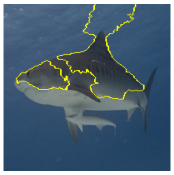

Figure 2: Shark

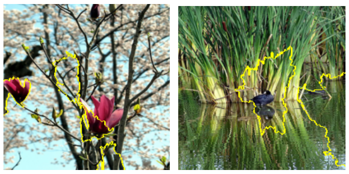

Figure 3: Kite and Coot

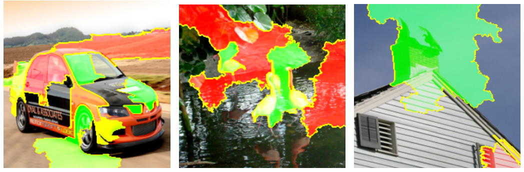

Figure 4: Areas contributing against top prediction

The annotations are obtained for the top 5 features of 10 image instances, considering two scenarios: annotating areas that encourage the top prediction and annotating regions of the image that contribute against the top prediction. Insights obtained from both scenarios are discussed in the report as follows. Figure 1 consists of a few annotated images from the first scenario. It can be observed that the model's predictions are highly influenced by prominent features like the white terrier's eyes, ears, and furry body. Additionally, the textures on common iguana and its eyes are annotated suggesting that these play a major role in the final prediction. Moreover, the shape and color of the flamingos are possibly responsible for the obtained annotation. A major part of the two oranges is also annotated, potentially because of the orange color. This explanation of the images can be attributed to their background being less complex in terms of colors, and other objects (of the same/different color as the predicted object). However, as we look at the images in Figure 3, it is clear that the model's predictions are influenced by some contextual components of the image too.

In the more complex image of the kite in Figure 3, along with the flower which is the main object of interest, other irrelevant features like branches also contribute to the prediction. Similarly, in the image of the American coot, it is prominent that along with the bird, the neighboring greenery and water are also captured suggesting that the model learns to associate the greenery and water plants to the habitat of the American coot, thus the annotation. This leads to another insight that the contextual information in the image also influences the model prediction.

Figure 2 is an image of two tiger sharks if we observe closely. An interesting insight to notice here is that the larger tiger shark's features like blunt nose and fins, are more dominant among the two also because they're more visible and apparent, so only the large shark's features are annotated, and the smaller shark is not annotated. Moreover, the American coot in Figure 3 covers a relatively smaller portion of the image, which can also be the potential reason why the annotation is not just limited to the bird's shape. Thus, it can observed that the size of the object of interest also affects the model's prediction. When scenario 2 is considered wherein we turn on the areas in the image that contribute against the top prediction, we observe that for flamingos (Figure 4 - image in the middle), the model associates with other classes in the image like leaves, land, and water and thus negatively contribute to the prediction of a flamingo. For the image of a vulture(Figure 4 - image towards the right), we see objects like windows and textures of the striped design of a house contributing negatively to the top prediction. Similarly, for the racer image(Figure 4 - image towards the left), features like headlights, wheels, the front part of the car, and the road (which provides a better context) contribute positively to the top prediction. At the same time, if observed closely, a person is sitting inside the car, so the model captures that as well along with some irrelevant background elements like sand. These features contribute negatively to the model's top prediction. The other visualizations for this scenario can be accessed here. To summarize the results, it can be concluded that LIME explains individual predictions of the model and based on the annotations, multiple factors like shapes, colors, textures, patterns, size of the objects, number of objects, context, resolution of images, background noise, complexity can influence the model's prediction.