Comparing LIME and Grad-GAM

This report gives an overview of the insights obtained on implementing Task 4 from Assignment 4. The main idea of the report is to compare two interpretability methods, LIME and Grad CAM, which differ in multiple aspects. The results for LIME and Grad-CAM on the 10 images provided can be found here. These two methods have been compared using the IoU metric and the implementation can be found here. All the LIME and Grad CAM masks can be accessed here. As shown in Table 1, the IoU scores vary significantly across different instances, with an average IoU of 0.0848.

Images	IoU score
West Highland white terrier	0.1270
American hoot	0.1166
Racer	0.0431
Flamingo	0.0689
Kite	0.0037
Goldfish	0.0234
Tiger Shark	0.1407
Vulture	0.0933
Common Iguana	0.0009
Orange	0.2299
Average IoU	0.0848

Table 1: IoU scores for different instances

LIME and Grad CAM are quite different in their approach. While LIME is more versatile and is perturbation based, Grad CAM is CNN specific and gradient based. Based on the scores in Table 1, it can be stated that for less complex images such as orange, white terrier, goldfish, tiger shark, american hoot, the IoU is relatively higher as compared to others. On the other hand, it significantly drops for more complex images like kite, vulture, flamingo, and common iguana. This behavior can be potentially attributed to LIME approximating a model locally around a prediction and Grad-CAM utilizes gradients to highlight important regions. For less complex images, the main features are more likely to be detectable by different interpretability methods which causes a significant overlap between their respective highlighted regions, which is not the case for complex images. The IoU scores for kite and common iguana are the lowest, probably because their different colors, textures, and patterns add to their complexity. This suggests that these two methods disagree on intricate and complex images.

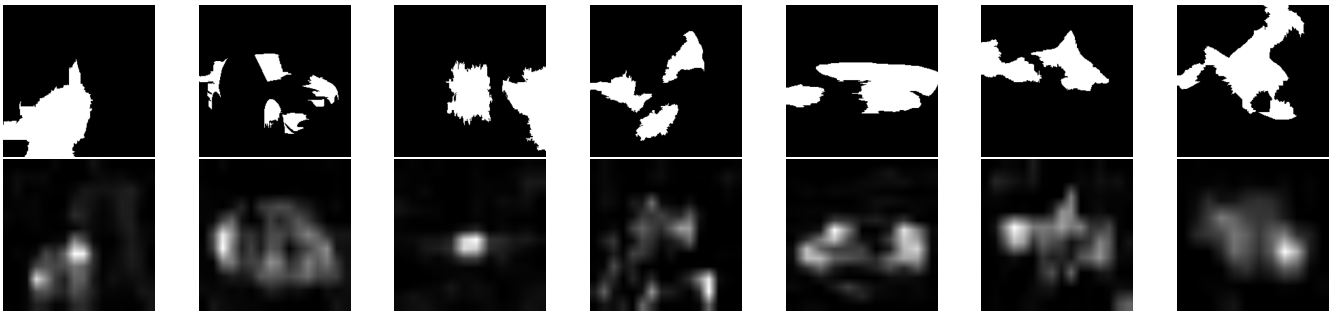

Figure 1: Upper Row - LIME masks, Lower Row - Grad CAM masks

Figure 1 consists of the masks generated by LIME and Grad CAM for white terrier, racer, American coot, common iguana, orange, tiger shark, and gold fish (from left to right). For almost all the ImageNet images provided, the masks obtained for LIME and Grad CAM methods look quite similar in terms of the activated regions and highlight important features. However, the Grad CAM masks are relatively very less sharp and smooth. Even after performing the experiments with aug smooth and eigen smooth, the obtained masks do not become as sharp as the ones obtained by LIME. While LIME annotates only specific important and distinct features responsible for the model's predictions, Grad CAM masks are less distinct and blurred but focus on the entire area where the object of interest is located inside the image. For the first image of the white terrier, the highlighted region for LIME is uniform and continuous but for Grad CAM, we see that the eyes and ears of the white terrier are specifically highlighted reflecting their greater importance over other features, as its other parts are blurred and have a diffused focus. If the second image of the racer is considered, a similar pattern is seen with wheels getting highlighted the most for Grad CAM. We can also see the segmented regions of the racer for LIME, but a less sharp and blurred heatmap that focuses on the broader concept of a racer for Grad CAM. The third image of the American coot has relatively similar masks, indicating that the focus is on the bird, but LIME also focuses on the context (like water bodies) in this case, unlike Grad CAM to decide the model's predictions. In the sixth image of a tiger shark, just like LIME, Grad CAM also highlights the head and fins of the shark specifically. Similarly from all the results, we find that LIME results in more sharp and interpretable individual features of the objects, whereas Grad CAM provides a general sense of the important regions in the image. The potential reason for the variation in the generated masks for both methods is that LIME operates on a perturbation-based approach and Grad CAM works on a gradient-based aggregation at the final convolutional layer.