TML Assignment 4: Explainability

Task 4 Report: A Quantitative Comparison of Grad-CAM and LIME

Objective

This report provides a comparative analysis of Grad-CAM and LIME. The core of this analysis is a quantitative "agreeability" score, calculated using the Intersection over Union (IoU) metric between the explanation masks generated by both methods. The goal is to investigate how this agreeability changes for images of varying complexity and thereby understand the fundamental differences between the two techniques.

Methodology: An IoU Metric for Agreeability

To create a quantitative comparison, we measured the IoU score directly between the explanation masks of the two methods. For each image, the process was:

1. **LIME Mask:** A binary explanation mask was generated using the complexity-tuned LIME approach from Task 3.
2. **Grad-CAM Mask:** The continuous heatmap from Grad-CAM was converted into a binary mask by applying a threshold. Only pixels with a normalized intensity value greater than **0.5** were included.
3. **IoU Calculation:** The IoU was then calculated between the LIME mask and the thresholded Grad-CAM mask. The resulting score represents the method's "agreeability", with a high score indicating strong overlap.

Quantitative Results

The following table summarizes the calculated agreeability IoU scores for all ten images. A higher score indicates stronger agreement between the LIME mask and the salient regions of the Grad-CAM heatmap.

Table 1: Summary of Agreeability IoU Scores between LIME and Grad-CAM.

Image Name	Agreeability IoU Score
West_Highland_white_terrier	0.274
racer	0.212
tiger_shark	0.222
common_iguana	0.207
goldfish	0.199
vulture	0.173
flamingo	0.117
orange	0.106
American_coot	0.079
kite	0.041

Comparative Analysis

The Effect of Image Composition on Method Agreement

The quantitative results in Table 1 strongly confirm that the agreeability between Grad-CAM and LIME is highly correlated with the clarity and composition of the image.

- **High Agreeability on Well-Defined Objects:**
  The `West_Highland_white_terrier` image produced the highest agreeability score (IoU: 0.274). This is because the subject is large, well-defined, and contiguous. Grad-CAM produces a strong, focused heatmap on the dog's face, and LIME's superpixel-based mask accurately captures this same region.

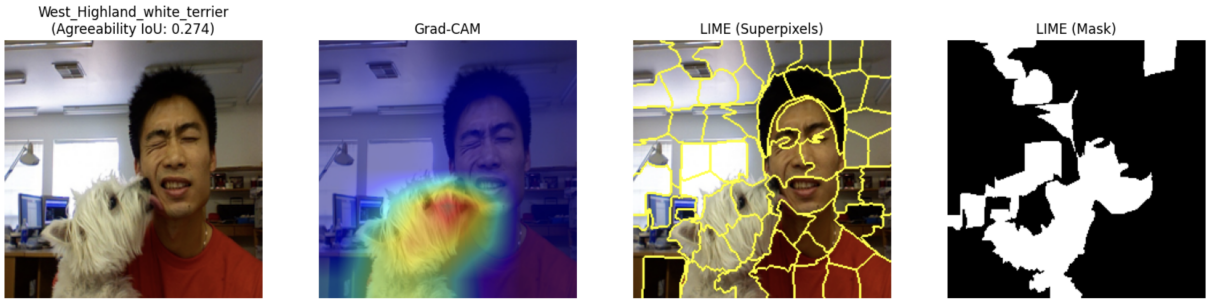

Figure 1: On the `terrier` image, the high IoU score shows strong agreement, as both methods successfully isolate the well-defined subject.

- **Low Agreeability on Complex or Misclassified Images:**
  The `kite` image produced the lowest IoU score (0.041). This quantitatively demonstrates the strong disagreement between the two methods when the model is confused. Grad-CAM produces a diffuse highlight on floral patterns, while LIME singles out a few specific superpixels it deemed most influential. The near-zero IoU is a direct measure of this disagreement. Similarly, the `American_coot` (IoU: 0.079) shows low agreement because the small, dark bird in a busy background causes the methods to highlight different, non-overlapping features.

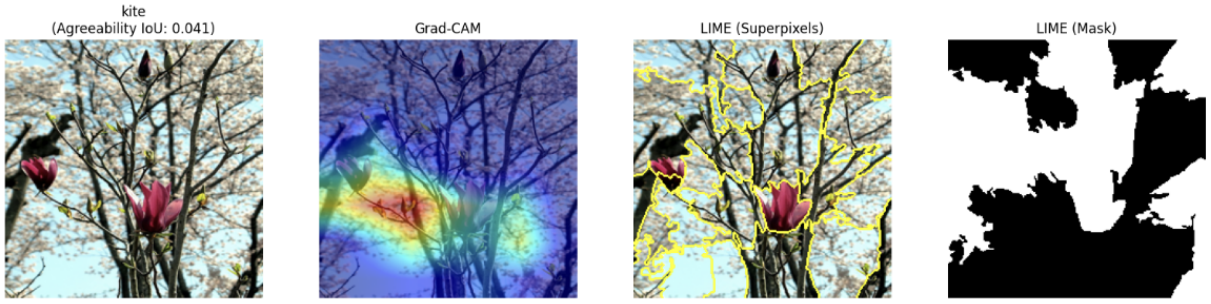

Figure 2: On the misclassified `kite` image, the extremely low IoU score confirms that the methods disagree on what caused the prediction.

Conclusion: Insights from Agreeability

Using a direct IoU agreeability score reveals key insights into the behavior of Grad-CAM and LIME.

1. **IoU as a Confidence and Clarity Proxy:**
   The agreeability IoU score is an effective proxy for both model confidence and image clarity. High IoU scores (above approx. 0.20) correlate with images where a single, clear object is present. Low scores correlate with model confusion or scenes where the subject is small, occluded or set against a cluttered background.

2. **Explaining Different Aspects of a Prediction:**
   The core reason for disagreement is that the methods explain different things. Grad-CAM shows a holistic, gradient-based attention map, a raw view of "where" the model is looking. LIME provides a more human-interpretable, parts-based explanation by identifying "which segments" were most influential in a local approximation.

In summary, while Grad-CAM shows a "soft" area of focus, LIME identifies "hard" influential components. Using IoU to measure their agreement is a powerful technique for quantitatively assessing when these two views of model behavior converge or diverge.