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. This score represents the method's "agreeability".

Quantitative Results Summary

The table below 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
tiger_shark	0.222
racer	0.212
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 of Key Examples

The results in Table 1 confirm that image composition directly impacts method agreement. We analyze three key cases below.

Highest Agreeability (`terrier`)

The `West_Highland_white_terrier` image produced the highest agreeability score (IoU: 0.274). This is because the subject is large, contiguous, and clearly distinct from the background. Grad-CAM produces a strong heatmap on the dog's face, and LIME's superpixel mask is clean. The high IoU score quantitatively confirms that for such clear subjects, both methods converge on a similar explanation.

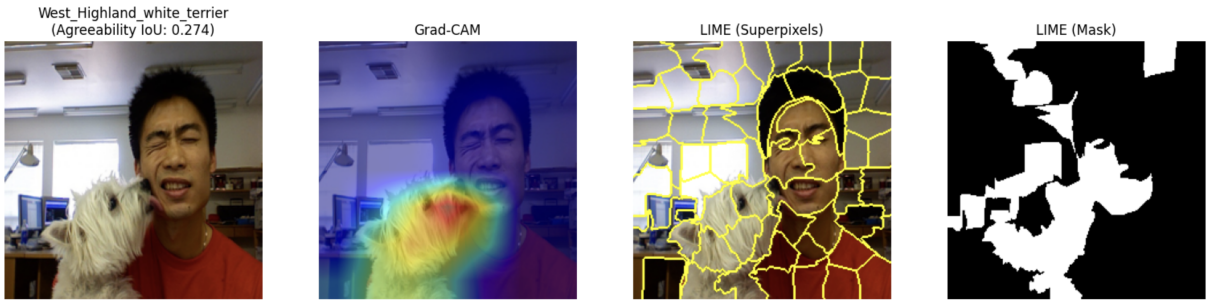

Figure 1: Highest agreeability on a clear, well-defined subject.

Simple vs. Complex Images

The assignment asks if IoU is higher for simple images like `goldfish` than for complex images like `kite`. Our results confirm this is emphatically the case.

The `goldfish` image features a clear subject against a simple background and produced a relatively high agreeability score of **IoU = 0.199**. As shown in Figure 2, the methods largely agree.

In stark contrast, the visually complex and misclassified `kite` image yielded the lowest score of the set at **IoU = 0.041**. This near-zero score, shown in Figure 3, reflects a

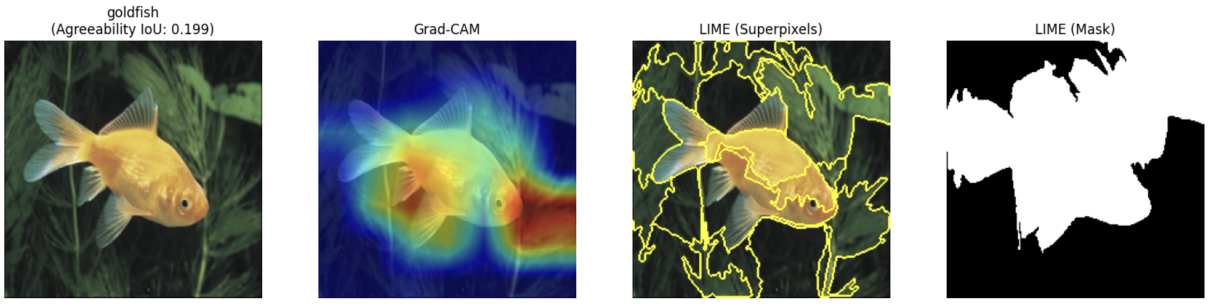

Figure 2: The `goldfish`: A simple image resulting in strong method agreement.

total disagreement between the methods, which expose different aspects of the model's confusion.

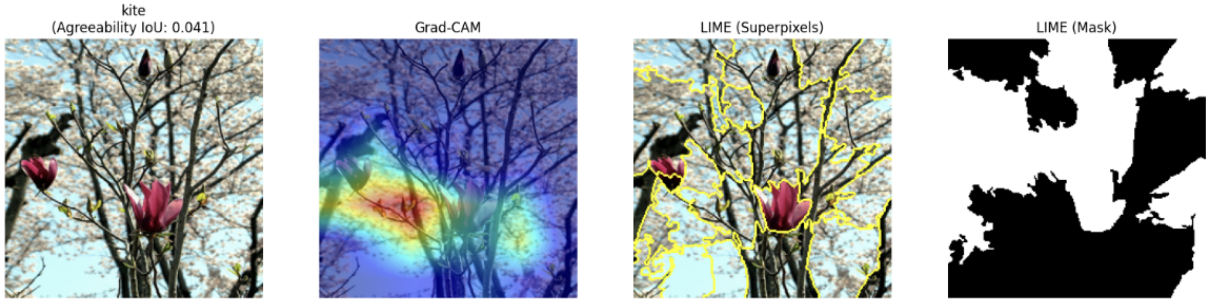

Figure 3: The `kite`: A complex image resulting in very low agreement.

Conclusion

By quantifying the agreement between Grad-CAM and LIME with an IoU score, we can draw clear conclusions.

1. **IoU as a Proxy for Clarity:** The agreeability score is an effective proxy for image clarity and model confidence. Simple images with clear subjects yield high IoU scores, while complex, cluttered, or misclassified images result in low scores.

2. **Fundamental Methodological Differences:** The IoU scores reveal the core difference between the methods. Grad-CAM gives a holistic, gradient-based view ("where the model looks"), while LIME gives a parts-based, local approximation ("which segments matter"). Their agreement measures the overlap between these two distinct explanatory views.