TML ASSIGNMENT 4

TASK 4

TEAM 16

AHRAR BIN ASLAM

MUHAMMAD MUBEEN SIDDIQUI

Objective

To analyze how two popular interpretability methods **Grad-CAM** and **LIME** differ in their visual explanations and investigate whether certain types of images lead to higher agreement between them, as measured by Intersection over Union (IoU).

Our Approach

The comparison of LIME and Grad-CAM explanations follows a structured approach. First, for LIME explanation generation (as done in Task 3 with the same parameters), each image is resized and passed through the LimeImageExplainer. A Quickshift segmentation algorithm is applied to generate superpixels, and a Ridge regression model is used to fit a local surrogate model. The top 10 most influential superpixels for the top predicted class are selected, resulting in a binary explanation mask that highlights the most important regions, which is then saved and visualized.

For Grad-CAM, pre-generated heatmaps from Task 2 are loaded for each image. These heatmaps are transformed into binary masks by thresholding the top 20% of intensity values, thereby highlighting the most activated regions in the image from the model's perspective.

To quantitatively compare both explanation methods, the Intersection over Union (IoU) is computed between the LIME and Grad-CAM binary masks for each image. Additionally, visualizations are created and saved for each sample, showing the original image, the Grad-CAM mask, the LIME mask, and their intersection. Finally, an average IoU score is computed across all images to assess the overall level of agreement between the two interpretability methods.

ANALYSIS:

The results of the IoU comparison reveal varying degrees of agreement between the LIME and Grad-CAM explanations across different images. The highest IoU was observed for the *tiger_shark* image (0.1971), indicating a relatively strong overlap in the regions highlighted by both methods. On the other hand, images like *orange* (0.0739) and *vulture* (0.0790) showed very low IoU scores, suggesting that LIME and Grad-CAM identified different regions as important.

Interestingly, the results do not show a clear pattern correlating image complexity with higher agreement. For instance, although the *goldfish* is a visually simpler image, its IoU

score was relatively low (0.0924), while more complex images like *kite* achieved slightly higher scores (0.1391). This suggests that the level of agreement may depend more on how each method interprets feature importance rather than on the visual complexity of the image.

Overall, the average IoU across all images was 0.1198, which indicates a modest level of overlap between the two methods. This reflects the fundamental differences in how LIME and Grad-CAM and LIME generate explanations. The visualizations further support these findings, showing that while some regions align, others differ significantly between the two approaches. These differences underscore the importance of using multiple interpretability methods to gain a more comprehensive understanding of model behavior.

RESULTS:

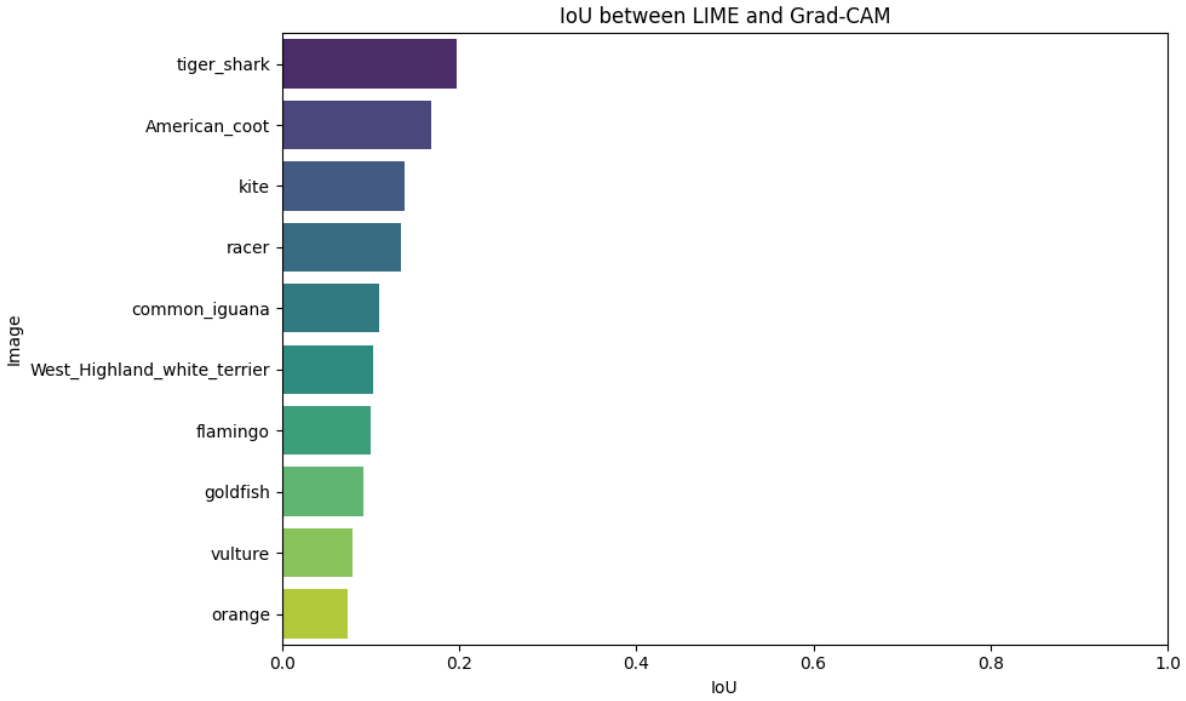
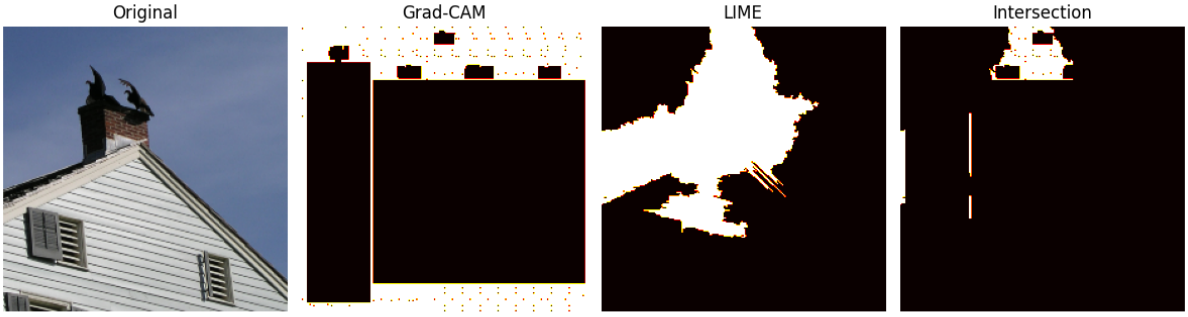
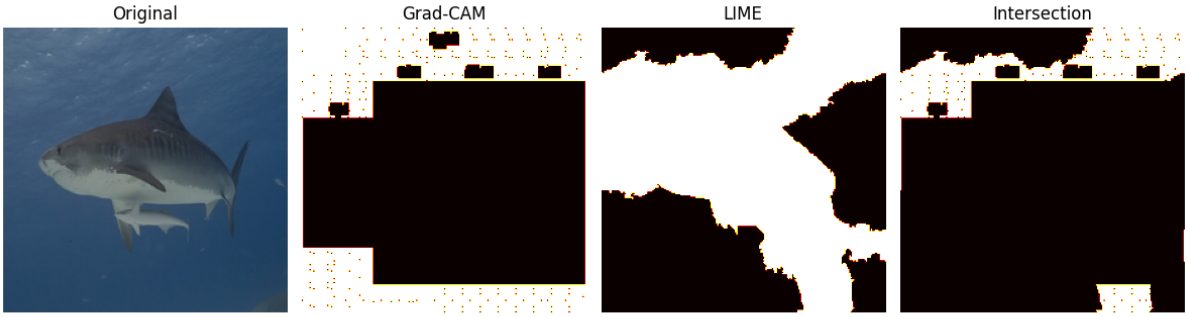
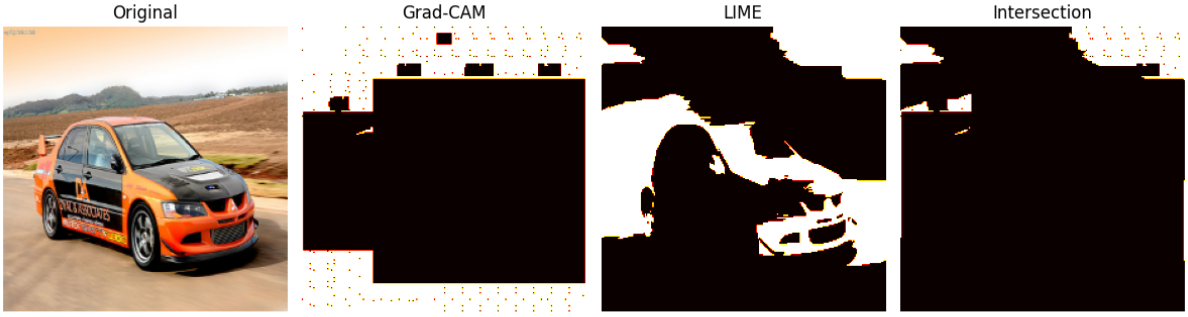


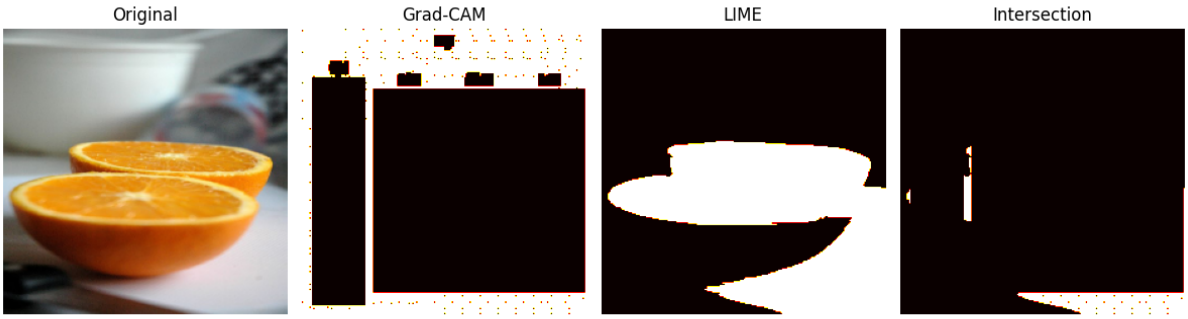
Image Class	IoU Score
West Highland White Terrier	0.1025
American Coot	0.1692
Racer	0.1343
Flamingo	0.1007
Kite	0.1391
Goldfish	0.0924
Tiger Shark	0.1971
Vulture	0.0790
Common Iguana	0.1101

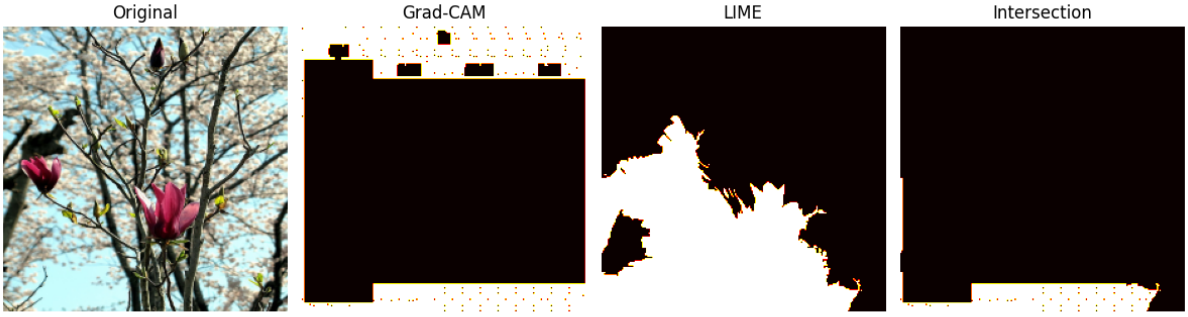
Average IoU: 0.1198

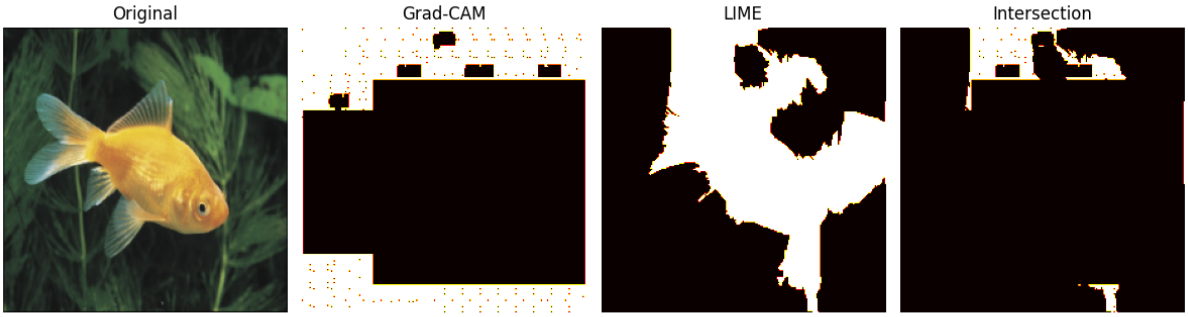

Original
Grad-CAM
LIME
Intersection

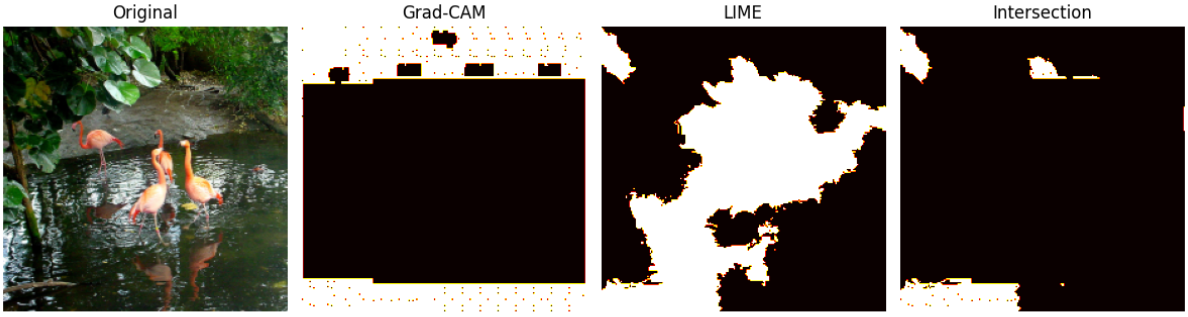

Original
Grad-CAM
LIME
Intersection

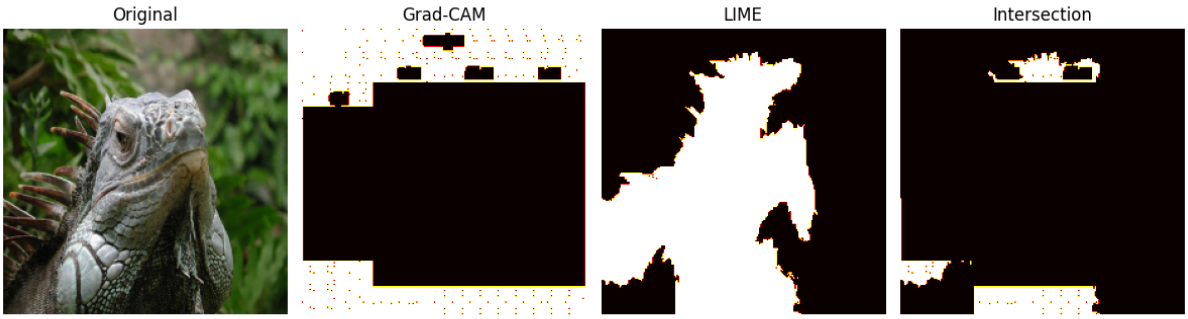

Original
Grad-CAM
LIME
Intersection

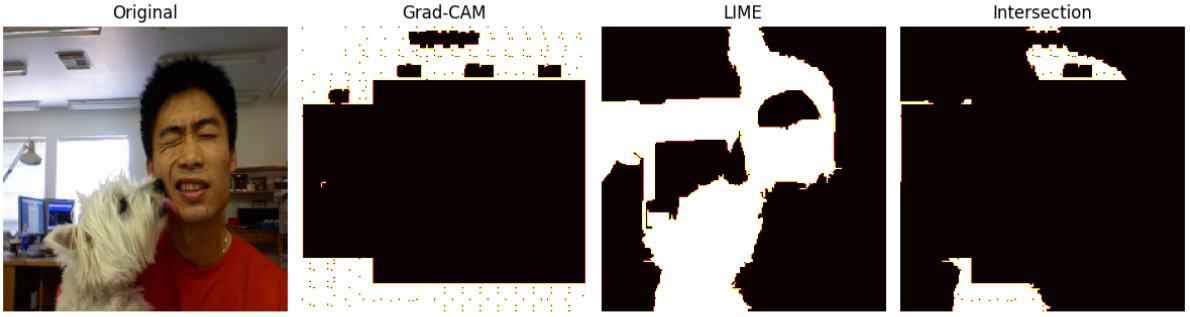

Original
Grad-CAM
LIME
Intersection

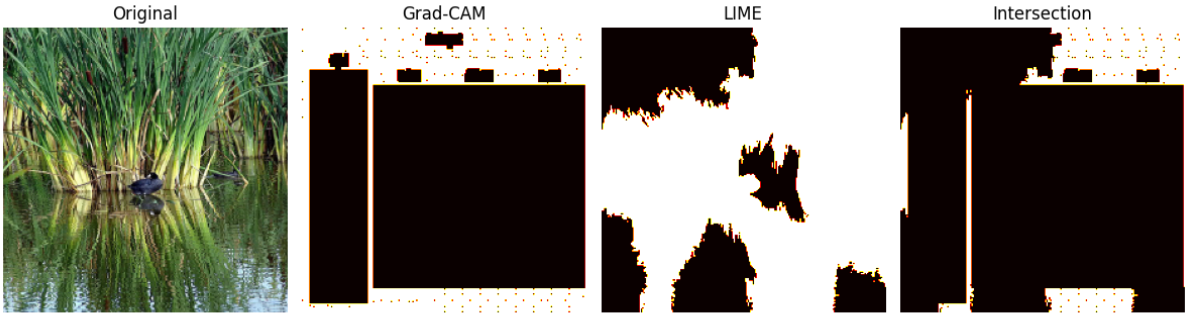

Original
Grad-CAM
LIME
Intersection


Original
Grad-CAM
LIME
Intersection


Original
Grad-CAM
LIME
Intersection


Original
Grad-CAM
LIME
Intersection


Original
Grad-CAM
LIME
Intersection


Original
Grad-CAM
LIME
Intersection