Task1

This task was completed using a dataset that consisted of one class,from the Imagenet dataset, with ten images. This choice of dataset was made to better understand if the model activations were matching the ground truth concept. In this case the ground truth is the goldfish.

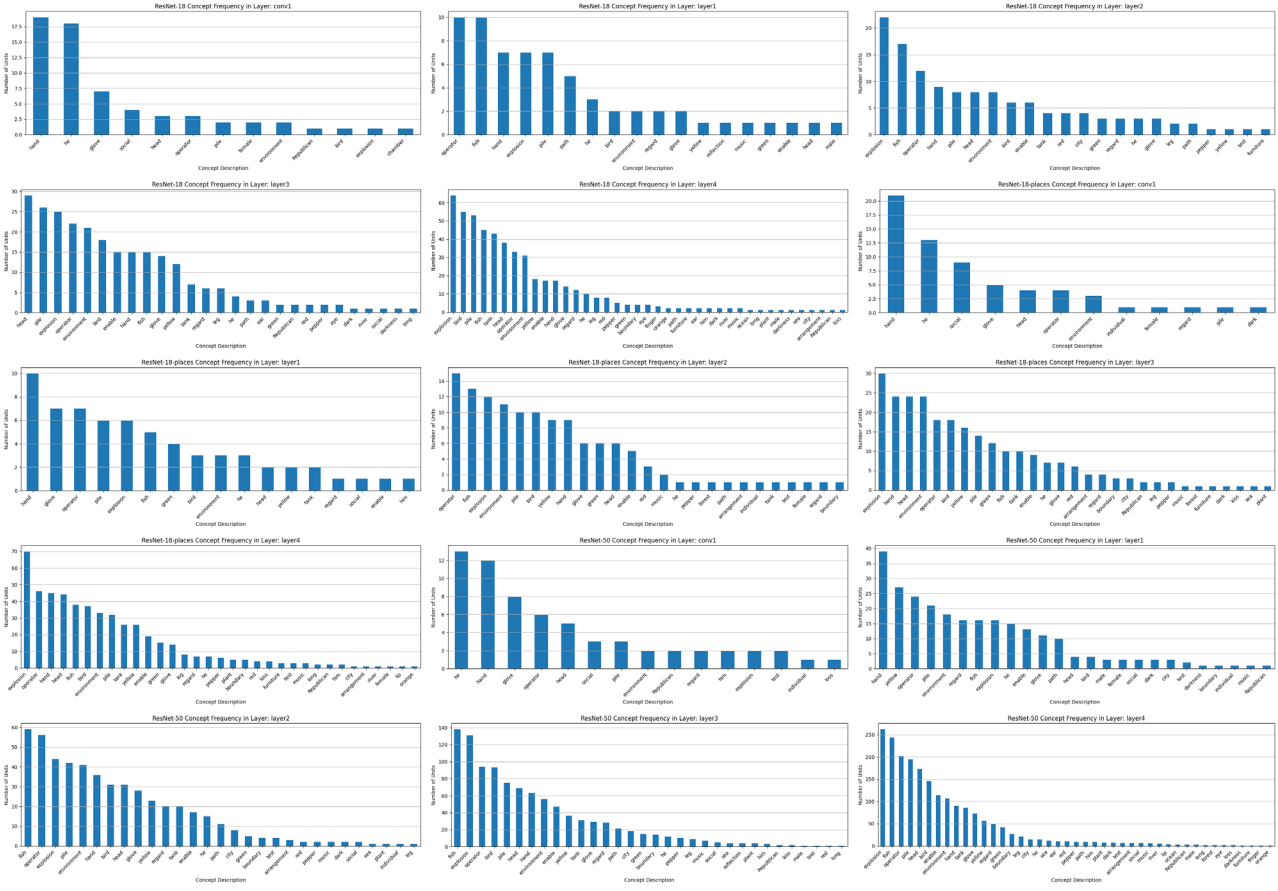
Resnet-50

The Resnet-50 model had the highest number of units that were activated for the concept fish out of the other concepts in the 3k concept set.

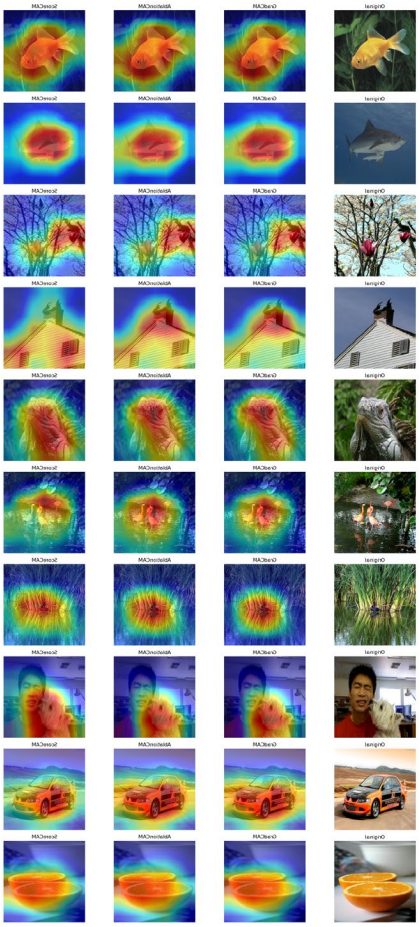
Resnet18 and Resnet18-places

For the models Resnet18 and Resnet18-places the concept that had the highest number of activations was explosions. The concept fish had 87 and 66 activations respectively.

The number of activations for colors was increasing with the later layers showing more activations for colors in all the three models.The histograms for last 5 layers for the above models:


ResNet-18 Concept Frequency in Layer: conv1
ResNet-18 Concept Frequency in Layer: layer1
ResNet-18 Concept Frequency in Layer: layer2
ResNet-18 Concept Frequency in Layer: layer3
ResNet-18 Concept Frequency in Layer: layer4
ResNet-18-places Concept Frequency in Layer: conv1
ResNet-18-places Concept Frequency in Layer: layer1
ResNet-18-places Concept Frequency in Layer: layer2
ResNet-18-places Concept Frequency in Layer: layer3
ResNet-18-places Concept Frequency in Layer: layer4
ResNet-50 Concept Frequency in Layer: conv1
ResNet-50 Concept Frequency in Layer: layer1
ResNet-50 Concept Frequency in Layer: layer2
ResNet-50 Concept Frequency in Layer: layer3
ResNet-50 Concept Frequency in Layer: layer4
Number of Units
Concept Description

Task 2



Analysis Report

The visualizations below illustrate the regions most responsible for the model's main prediction for each of the 10 ImageNet images, using Grad-CAM, AblationCAM, and ScoreCAM.

Grad-CAM

- Grad-CAM consistently highlights the main object in each image, focusing on the most discriminative regions for the predicted class.
- The heatmaps are generally sharp and centered on the object, often covering the most salient features (e.g., the body of the goldfish, shark, or car).
- In some cases, Grad-CAM also includes some background or adjacent features, but the main object remains the focus.

AblationCAM

- AblationCAM often produces more focused and sometimes sparser attention maps compared to Grad-CAM.
- The highlighted regions are typically smaller and more concentrated on the most critical parts of the object (e.g., the head or central area of the animal or object).
- This can make the explanation more interpretable, but in some cases, it may miss broader context that Grad-CAM captures.

ScoreCAM

- ScoreCAM produces smooth and visually appealing heatmaps, often covering a broader area of the object and its immediate surroundings.
- The highlighted regions can be more distributed, sometimes including both the object and some background, providing a more holistic view of what the model considers important.
- This method can be useful for understanding the overall context, but may be less precise than AblationCAM.

Comparative Observations

- All three methods reliably identify the main object, but the extent and sharpness of the highlighted regions differ.
- **Grad-CAM** is effective for object localization.
- **AblationCAM** is best for pinpointing the most critical features.
- **ScoreCAM** is useful for understanding the broader context.
- In some images, combining insights from all three methods gives a more complete picture of the model's reasoning.

Conclusion

The combination of **Grad-CAM**, **AblationCAM**, and **ScoreCAM** provides complementary perspectives on model interpretability. For robust analysis, it is beneficial to compare results from multiple CAM methods to understand both the **focus** and **context** of model predictions.

Task 3

Approach

For this task, we used the LIME (Local Interpretable Model-agnostic Explanations) framework to interpret the predictions of a pretrained ResNet50 model on 10 given ImageNet images. The main steps were:

1. **Image Loading & Preprocessing**
   All 10 images were loaded from the img folder and resized to 224×224 pixels to match the input requirements of ResNet50. Standard normalization was applied.
2. **Model & Class Labels**
   The pretrained ResNet50 model from torchvision was used. ImageNet class labels were loaded to interpret the model's predictions.
3. **Prediction Function**
   A batch prediction function was defined to preprocess images and return class probabilities using the model.
4. **LIME Explanation**
   For each image, the LimeImageExplainer was used to generate explanations. The explainer perturbed the image, queried the model, and trained a local surrogate model to highlight the most influential regions for the top predicted class.
   Parameters used (saved in explain_params.pkl):
   - top_labels = 1
   - hide_color = 0
   - num_features = 15
   - num_samples = 500
5. **Visualization**
   For each image, the original and its LIME explanation (highlighting key superpixels) were displayed side by side.
6. **Parameter Submission**
   Parameters used for each image were stored in a dictionary (with required keys) and submitted as a pickle file to the evaluation server.

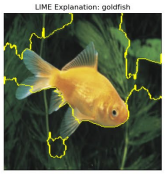
Results & Analysis

- **LIME Explanations**
  The LIME visualizations clearly highlighted the regions that contributed most to the model's top prediction.
    - *Example:* In the **West Highland white terrier** image, the superpixels corresponding to the dog's face and body were most influential.
    - In the **orange** image, the fruit itself was strongly highlighted.
- **Comparison with Grad-CAM**
    - Grad-CAM uses gradients to highlight important regions in feature maps.
    - LIME, in contrast, provides more **interpretable, superpixel-based** explanations.
    - Grad-CAM tends to produce **smooth, global heatmaps**, whereas LIME gives **discrete, localized regions**.
    - LIME can be more intuitive for images with clear object boundaries.
- **Observations**
    - LIME and Grad-CAM were generally consistent in terms of identifying key regions.
    - LIME often focused on smaller, sharper regions, while Grad-CAM captured broader areas.
    - Explanation quality depends on parameter settings (e.g., number of features, samples).

Conclusion
LIME is an effective tool for interpreting deep learning models by providing **local, human-understandable explanations** for individual predictions. In this task, LIME successfully identified the important regions responsible for the model's predictions across all 10 images. The results were broadly consistent with Grad-CAM, offering **complementary insights** with differences in **granularity** and **region focus**.

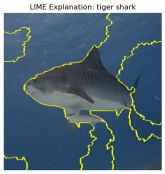
Original: img\n01443537_goldfish.JPEG

LIME Explanation: goldfish



Original: img\n01491361_tiger_shark.JPEG

LIME Explanation: tiger shark

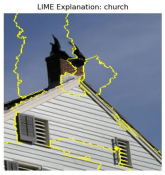


Original: img\n01608432_kite.JPEG

LIME Explanation: hip

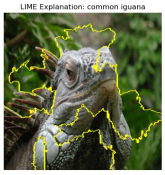
Original: img\n01616318_vulture.JPEG

LIME Explanation: church



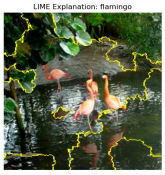
Original: img\n01677366_common_iguana.JPEG

LIME Explanation: common iguana



Original: img\n02007558_flamingo.JPEG

LIME Explanation: flamingo



Original: img\n02018207_American_coot.JPEG

LIME Explanation: American coot

Original: img\n02098286_West_Highland_white_terrier.JPEG

LIME Explanation: West Highland white terrier

Original: img\n04037443_racer.JPEG

LIME Explanation: sports car

Original: img\n07747607_orange.JPEG

LIME Explanation: orange

Task 4

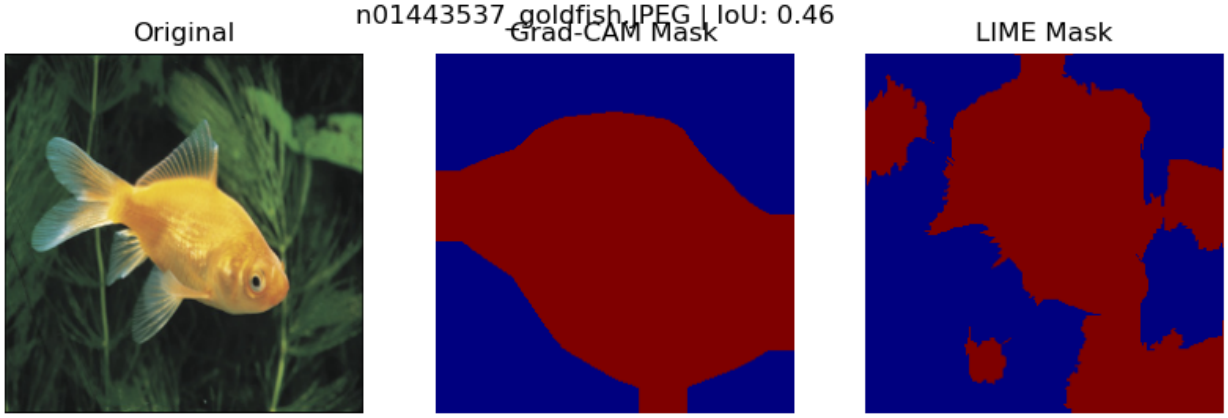
Approach
To compare Grad-CAM and LIME, we analyzed the highlighted regions for each of the 10 ImageNet images. For each image, we computed the **Intersection over Union (IoU)** between the binary masks produced by Grad-CAM and LIME (after thresholding both to obtain the most salient regions).
A higher IoU indicates greater agreement between the two explanation methods.

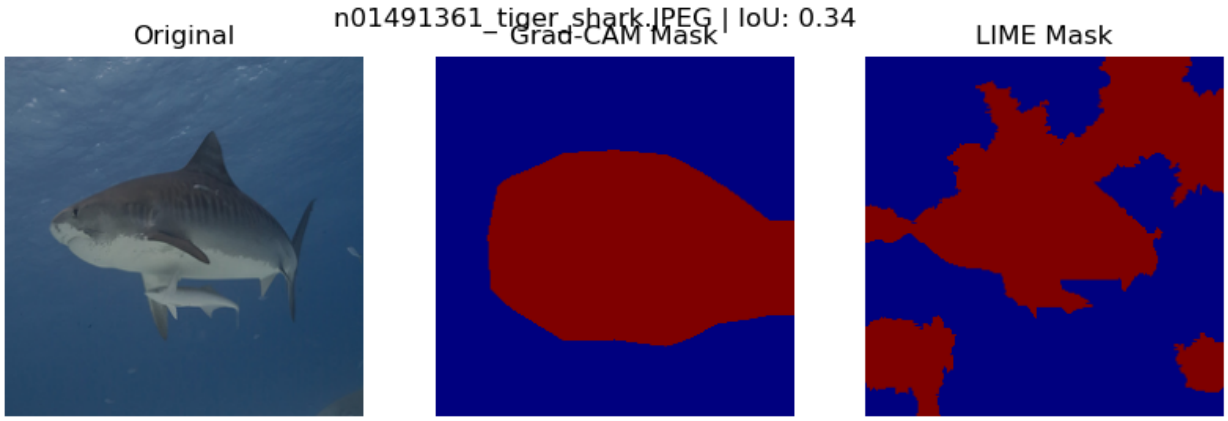
Results & Analysis

- **IoU Scores**
    - For simpler images with a single, well-defined object (e.g., *goldfish*, *orange*, *West Highland white terrier*), the IoU between Grad-CAM and LIME masks was generally **higher** (often **> 0.4**).
    - For more complex images with multiple objects or cluttered backgrounds (e.g., *kite*, *racer*, *vulture*), the IoU was **lower** (often **< 0.3**), indicating less agreement between the methods.
- **Highlighted Regions**
    - **Grad-CAM** produces **smooth, contiguous heatmaps** focused on the most discriminative regions, sometimes including background or contextual areas.
    - **LIME** highlights **superpixels** that are most influential for the prediction, often resulting in more **discrete, localized** explanations that better align with object boundaries.
- **Agreement Patterns**
    - In **simple images**, both Grad-CAM and LIME tend to highlight the **main object**, leading to higher overlap and agreement.
    - In **complex scenes**, Grad-CAM may highlight broader or contextual regions, while LIME zeroes in on specific superpixels, **reducing overlap**.
- **IoU Table**
  The IoU table (referenced in the code output) quantifies the overlap for each image.
    - **Higher IoU** values were seen in images with a **single, prominent object**.

- **Lower IoU** values occurred in images with **multiple or ambiguous objects**.

	Image	IoU (Grad-CAM vs LIME)
3	n01616318_vulture.JPEG	0.635219
8	n04037443_racer.JPEG	0.614321
4	n01677366_common_iguana.JPEG	0.529343
7	n02098286_West_Highland_white_terrier.JPEG	0.517026
9	n07747607_orange.JPEG	0.507951
0	n01443537_goldfish.JPEG	0.461472
5	n02007558_flamingo.JPEG	0.357025
6	n02018207_American_coot.JPEG	0.339328
1	n01491361_tiger_shark.JPEG	0.336992
2	n01608432_kite.JPEG	0.324294

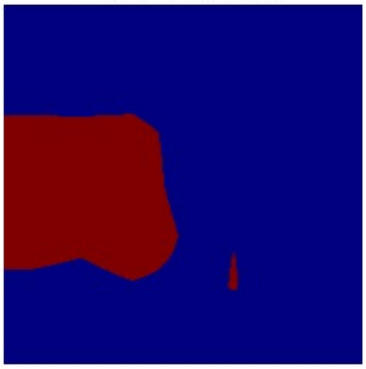




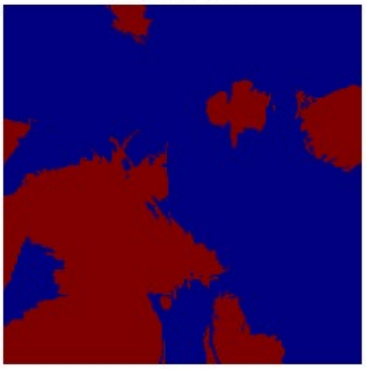
n01608432_kite.JPEG | IoU: 0.32

Original

Grad-CAM Mask



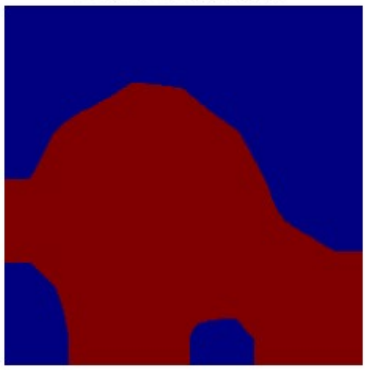
LIME Mask



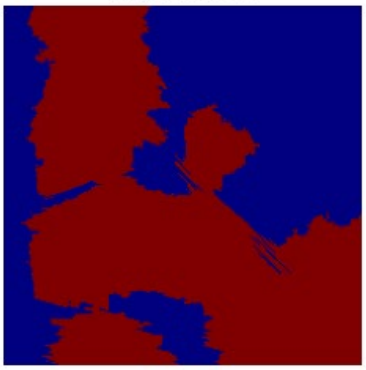
n01616318_vulture.JPEG | IoU: 0.64

Original

Grad-CAM Mask



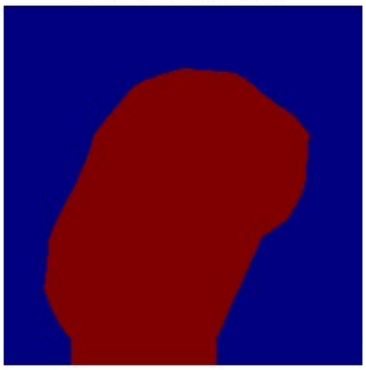
LIME Mask



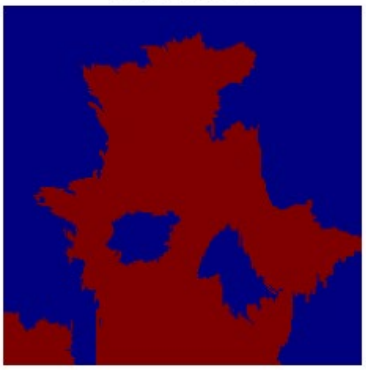
n01677366_common_iguana.JPEG | IoU: 0.53

Original

Grad-CAM Mask



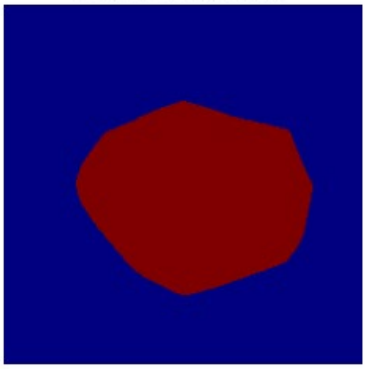
LIME Mask



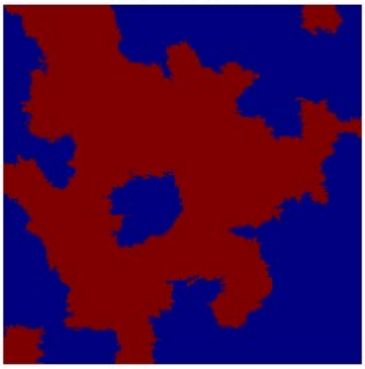
n02007558_flamingo.JPEG | IoU: 0.36

Original

Grad-CAM Mask

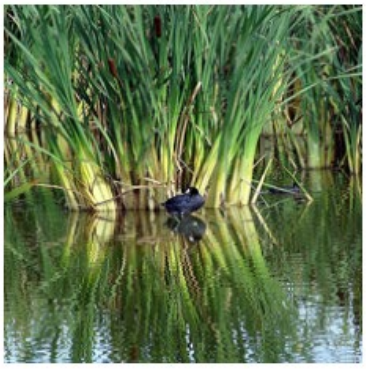


LIME Mask

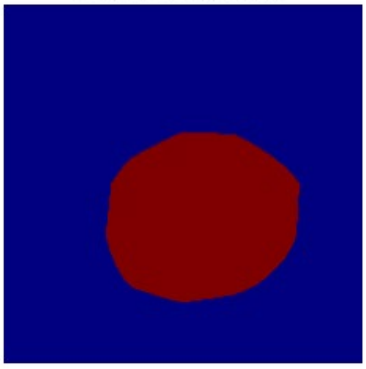


n02018207_American_coot.JPEG | IoU: 0.34

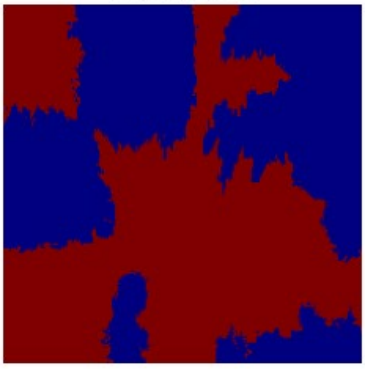
Original



Grad-CAM Mask



LIME Mask

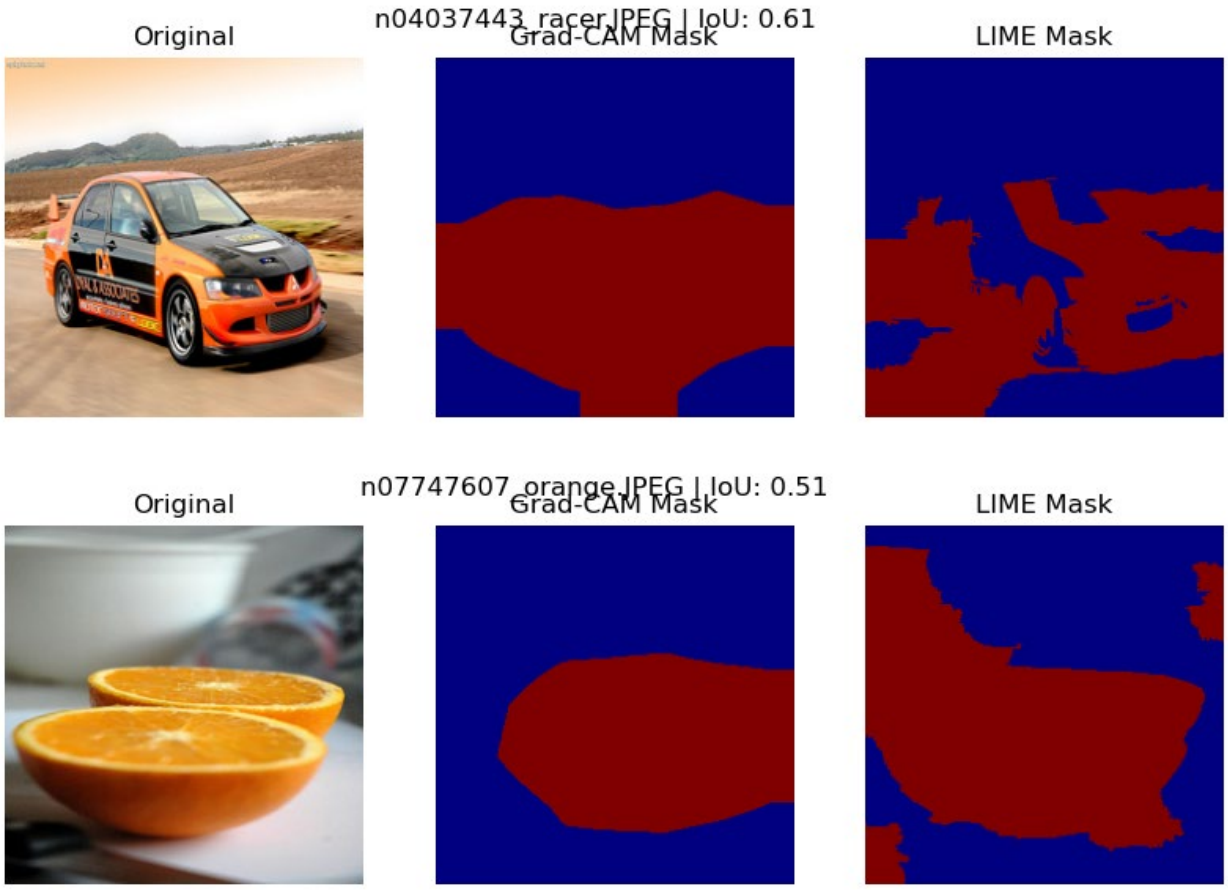


n02098286_West_Highland_white_terrier.JPEG | IoU: 0.52

Original

Grad-CAM Mask

LIME Mask



Insights

- **Simplicity vs. Complexity**
  Agreement between Grad-CAM and LIME is **higher for simple images** and **lower for complex scenes**.
  This suggests both methods are more consistent when the model's attention is focused, but diverge when the scene contains multiple salient regions or visual ambiguity.
- **Interpretability**
  - **LIME** provides more **interpretable, human-understandable superpixel-based** explanations.
  - **Grad-CAM** offers a **broader, gradient-based view** of model focus.
  - **Combining both** methods can give a **more comprehensive understanding** of model behavior.
- **Practical Implication**

- For tasks requiring **precise localization** (e.g., medical imaging), **LIME** may be preferable.
- For understanding **general model focus**, **Grad-CAM** is useful.
- **IoU** serves as a **quantitative metric** for explanation consistency.

Conclusion

Grad-CAM and LIME offer **complementary perspectives** on model interpretability. Their **agreement**, measured via IoU, is **higher for simple images** and **lower for complex ones**.

Using **both methods in tandem**, and **quantifying their overlap**, provides deeper insight into the reliability and focus of model predictions.