TML ASSIGNMENT 4

TASK 3

TEAM 16

AHRAR BIN ASLAM

MUHAMMAD MUBEEN SIDDIQUI

Objective

To generate interpretable explanations for predictions made by a pretrained ResNet-50 on 10 diverse ImageNet images using LIME (Local Interpretable Model-agnostic Explanations). The goal was to visualize which parts of each image contributed most to the model's top-1 classification decision.

Our Approach

In our approach, we applied LIME to 10 diverse ImageNet images using a pretrained ResNet-50 model. The objective was to visualize which image regions contributed most to the model's top-1 prediction, while optimizing for higher Intersection over Union (IoU) and lower execution time. Our implementation was carefully designed to balance both goals.

To begin with, ResNet-50 was selected for its strong ImageNet performance and its widespread use in computer vision applications. For preprocessing, two pathways were used: normalized tensors for model inference, and unnormalized float arrays for LIME's image perturbations. This separation-maintained model accuracy while ensuring LIME compatibility.

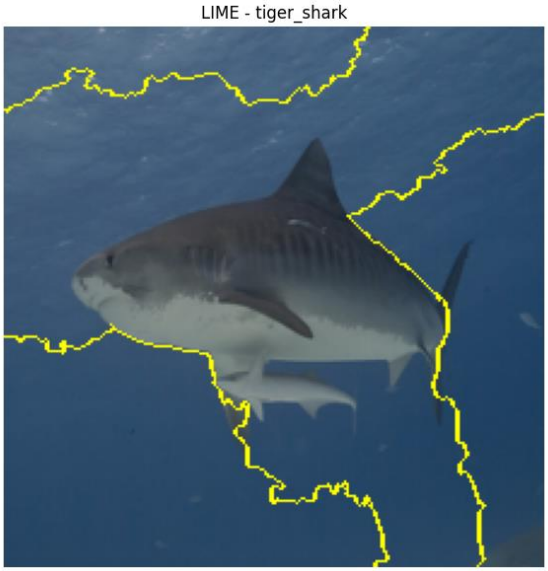
One of the key design choices was the use of Quickshift segmentation, which is edge-aware and better suited for natural images than basic grid or SLIC methods. This directly supports a higher Intersection-over-Union (IoU) between the LIME mask and object region, as the segmentation respects image boundaries more naturally. Furthermore, we used a 20th percentile pixel value as the occlusion color, instead of a fixed black or gray. This made the perturbations more realistic and preserved data distribution consistency, enhancing explanation fidelity and interpretability.

To keep the execution time low, we configured LIME to use only 700 perturbed samples and to focus on 12 most informative superpixels. This limits computational overhead without sacrificing much interpretability. A regularized Ridge regression model ($\alpha = 0.5$) was chosen as the surrogate model to ensure stability and prevent overfitting to noisy perturbations.

RESULTS

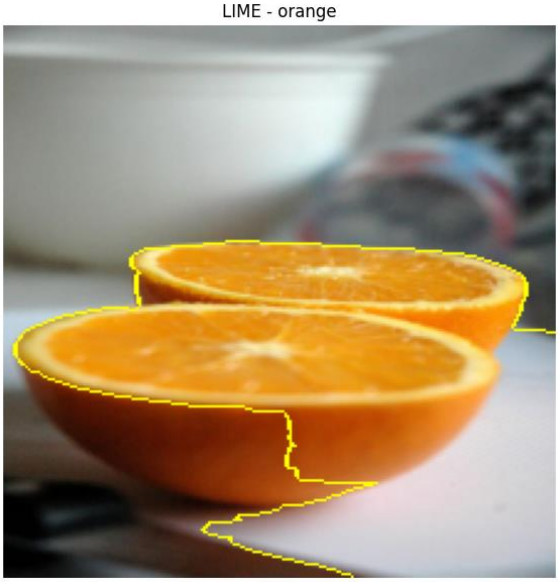
LIME - vulture

LIME - tiger_shark



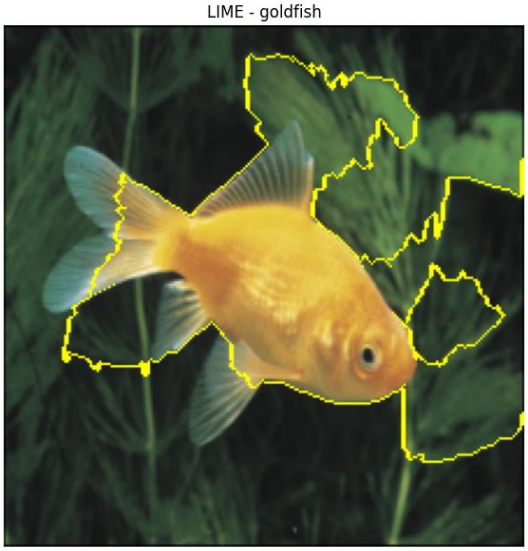
LIME - racer

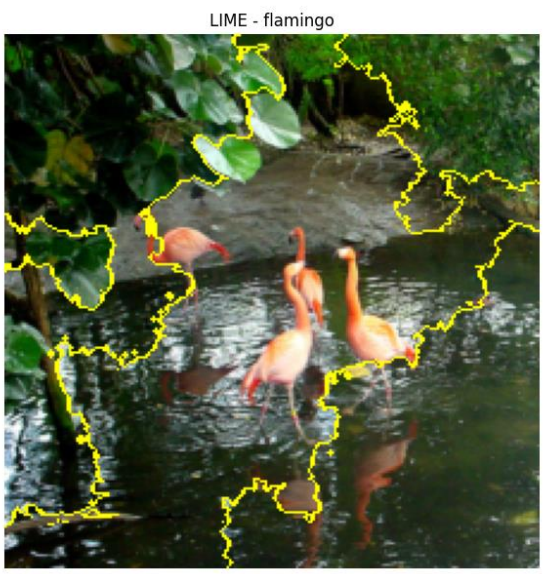
LIME - orange

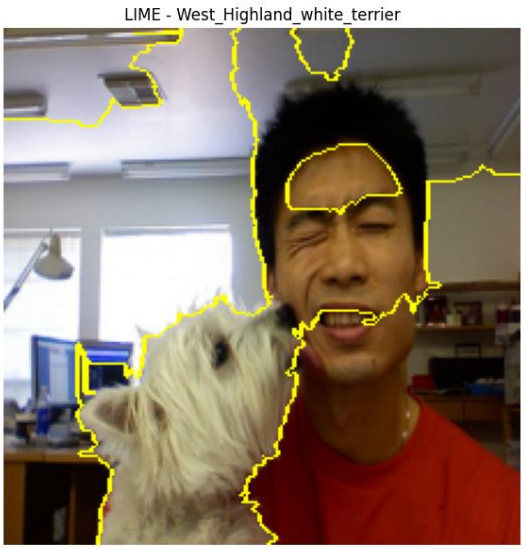


LIME - kite

LIME - goldfish







The explanations provided by LIME were both effective and efficient. When the object in the image was large, centered, and easy to distinguish, LIME did a great job of highlighting the most important parts. For example:

- In the **tiger shark** and **goldfish** images, LIME successfully focused on the body and fins of the animals, with very little interference from the background.
- In the **orange** image, LIME perfectly captured the circular shape of the fruit, showing how well it can isolate objects with distinct textures and colors.
- For the **iguana**, LIME highlighted the animal's head and spiny texture, which aligns with the key features the model likely used for classification.

However, LIME's performance wasn't as strong when the objects were smaller or when they blended with the background.

For example:

- In the **American coot** image, the bird was almost hidden among the reeds and reflections in the water. LIME struggled to focus only on the bird, leading to a lower overlap with the actual object.
- In the **kite** image, the model was likely confused by the similarity between the flowers and a flying kite, so LIME highlighted the flowers instead of the kite.
- For the **West Highland white terrier** image, both the dog and a person's face were marked as important, suggesting the model may rely on context, which LIME exposed.