Deliverable 05

Both LIME and Grad-CAM are widely used for model interpretability, but their explanations differ fundamentally:

- **LIME** works by perturbing superpixels and training a local surrogate model, resulting in sparse, sharply defined regions directly linked to pixels.
- **Grad-CAM** uses model gradients and feature maps to produce a smooth heatmap, showing where the model "looks" in a broader, less discrete way.

Simple Images: High Agreement

For images with a single, clear object and little background clutter such as **goldfish**, **tiger shark**, or **orange** LIME and Grad-CAM mostly agree. LIME explanations highlight the goldfish's body or the shark's outline, while Grad-CAM's heatmap also concentrates on these regions, particularly the head or main body. The calculated IoU (overlap of the explanation masks) for these images is generally higher, indicating strong agreement between the two methods.

Images with Multiple or Overlapping Objects

In scenes like **flamingo** or **American coot** (with several birds or some background complexity), both methods remain relatively focused, but some differences emerge. LIME tends to select each object individually, while Grad-CAM's heatmap sometimes merges them into one contiguous "hot" area, or slightly includes background reflections or nearby reeds. The IoU remains moderate but starts to drop compared to the simple-object cases.

Complex or Ambiguous Images: Lower Agreement

For images like **kite** or **vulture**, where objects are thin, fragmented, or not easily separated from the background, the explanations diverge more. Grad-CAM may highlight only the most discriminative region (such as the main bird in "kite" or the vulture's head), while LIME's superpixels might be fragmented across both object and background due to segmentation limits. This results in a lower IoU indicating the methods disagree more on what's important when the scene is visually complex or ambiguous.

Structured/Man-made Objects

In images like **racer**, both methods perform well, but Grad-CAM's heatmap spreads over the entire car while LIME targets specific car features (body, wheels, decals). The overlap is good but not perfect; Grad-CAM can be more inclusive, whereas LIME remains selective.

Observations

- **Agreement Depends on Image Simplicity:** For simple, single-object images, the two methods tend to agree, and IoU is high. In your set, images like goldfish, orange, and tiger shark exemplify this.
- **Complexity Reduces Agreement:** In cluttered scenes or with objects that blend into the background (e.g., kite, vulture), the two methods often disagree, with LIME's results more fragmented and Grad-CAM's focus more localized. IoU drops in these scenarios.
- **Nature of Explanation:** Grad-CAM often highlights broader regions, sometimes including some background, while LIME tends to be more precise but can be sensitive to segmentation and local pixel changes.
- **Interpretability:** LIME's explanations are easier to interpret pixel-by-pixel but can miss context. Grad-CAM provides a higher-level "attention map," which is smoother and sometimes less specific.