TML Assignment 4: Explainability

Task 2 Report: Grad-CAM Method Analysis

Objective

The purpose of this task was to analyze the predictions of a pre-trained `ResNet50` model using three Class Activation Map (CAM) techniques: Grad-CAM, AblationCAM, and ScoreCAM. By applying these methods to ten distinct images from the ImageNet dataset, we aimed to visualize the specific pixel regions the model considers most important for its predictions and to compare the characteristics of each explainability method.

Analysis of Visualizations

The analysis of the ten images reveals several key patterns regarding the model's focus and the behavior of the CAM methods. Overall, all three methods were successful in localizing the primary objects, confirming the model's predictions are based on relevant features. The main differences between the methods relate to the precision and sharpness of the generated heatmaps.

Comparison of CAM Methods

A clear hierarchy of precision was observed among the techniques:

- **Grad-CAM** consistently produced the most diffuse heatmaps. While correctly identifying the object region, it often included significant portions of the background.
- **AblationCAM** offered a noticeable improvement, generating heatmaps that were more focused on the subject than Grad-CAM.
- **ScoreCAM** reliably produced the cleanest and most tightly-focused visualizations. In nearly every case, its activation map was the most concentrated on the object's most discriminative parts, making its explanations the easiest to interpret.

Analysis of Representative Examples

To illustrate these findings, three representative images were chosen.

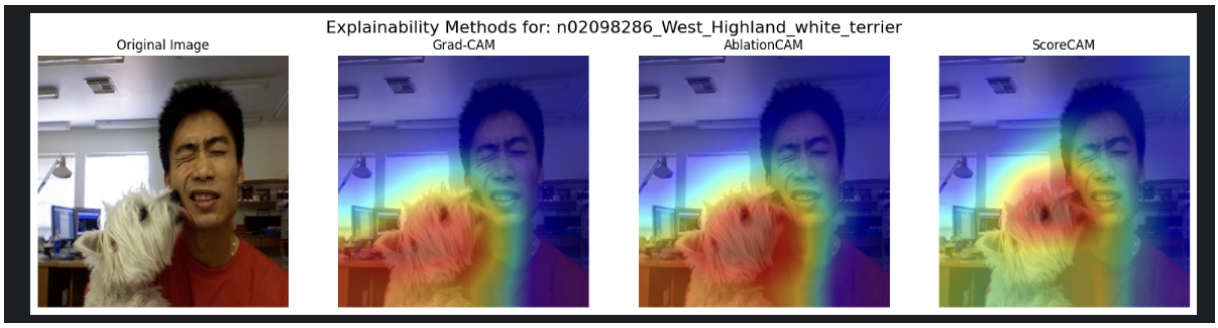

Figure 1: CAM results for the `West_Highland_white_terrier` image. All methods correctly focus on the terrier, ignoring the human face, with ScoreCAM providing the sharpest localization.

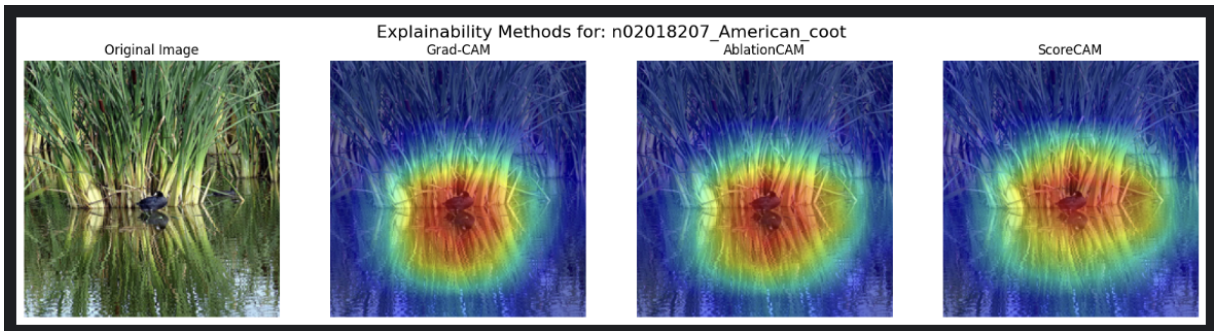

Figure 2: CAM results for the `American_coot`. The methods successfully pinpoint the small bird within a visually complex background of reeds, demonstrating robustness.

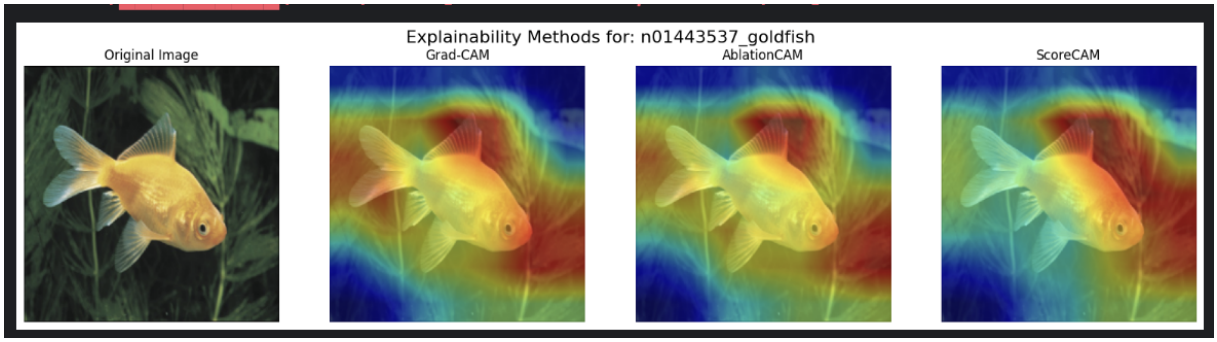

Figure 3: CAM results for the `goldfish`. In this simple case with a clear subject, all methods show high agreement, producing sharp, well-defined heatmaps on the fish's body.

Conclusion

The analysis across the ten images demonstrates the effectiveness of CAM-based methods for model explainability. Three main conclusions can be drawn:

1. **Model Reliability:** The `ResNet50` model is generally reliable, focusing its attention on semantically relevant objects when making predictions.

2. **Method Precision:** There is a clear trade-off among CAM methods. While all are effective, ScoreCAM provides the highest precision, generating sharper heatmaps that are ideal for localizing specific object features.

3. **Diagnostic Value:** These tools are valuable for confirming that a model is working as intended. The visualizations consistently show the model is not relying on spurious background correlations but on the features of the objects themselves.