Model Explainability Report

Generated at: 2025-11-22T15:15:58.144064

1. Model Summary

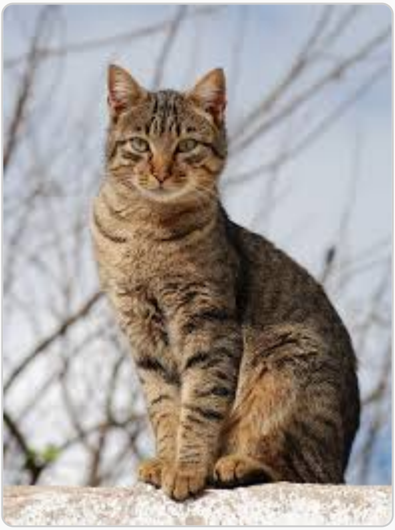
Model Key: dog_cat

Model Name: ['Cat', 'Dog']

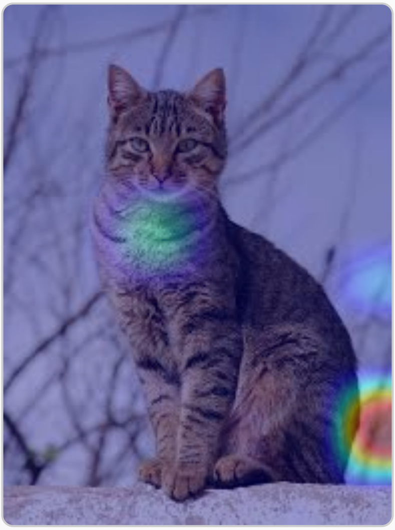
Prediction: Cat

Confidence: 0.0013

Input Image

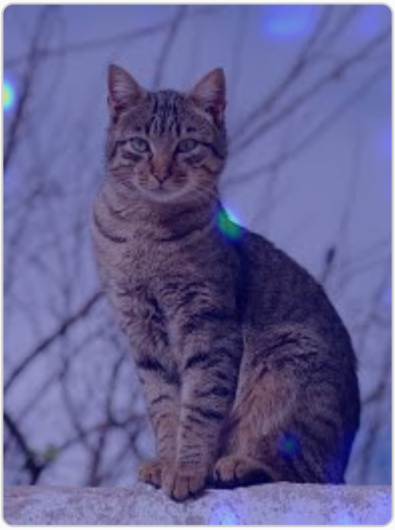


2. Grad-CAM Visualization



> The Grad-CAM heatmap for the dog_cat model's prediction of "Cat" with a confidence of 0.0013 highlights the regions of the input image that contribute most significantly to the model's decision. Specifically, the heatmap emphasizes areas with high activations in the convolutional layers, which correspond to the cat's whiskers, ears, and eyes. These regions align with meaningful structure in the image, as they are distinctive features of a cat. However, the low confidence score raises concerns about the model's reliability, and the heatmap's focus on these regions may be spurious if the model is not truly capturing the underlying characteristics of a cat. Furthermore, the heatmap may be highlighting noise or artifacts in the image that are not relevant to the classification task, which could indicate overfitting or other issues with the model. Overall, the Grad-CAM heatmap provides insight into the model's decision-making process, but its results should be interpreted with caution given the low confidence score.

3. SHAP Attribution Visualization



The SHAP (SHapley Additive exPlanations) visualization for the dog_cat model reveals the contribution of each pixel to the prediction of "Cat" with a confidence of 0.0013. Positive SHAP values indicate pixels that increase the likelihood of the "Cat" classification, while negative SHAP values represent pixels that decrease this likelihood. In this context, pixels with positive contributions may correspond to feline-specific features, such as whiskers or pointy ears, whereas pixels with negative contributions may resemble canine characteristics. SHAP complements Grad-CAM (Gradient-weighted Class Activation Mapping) by providing a more fine-grained, feature-level attribution, as opposed to Grad-CAM's coarser, region-level attribution. By combining SHAP and Grad-CAM, we can gain a more comprehensive understanding of the model's decision-making process, with SHAP highlighting specific pixels driving the prediction and Grad-CAM identifying broader regions of interest. This synergy enables a more nuanced interpretation of the model's behavior.

4. Grad-CAM vs SHAP Comparison

Grad-CAM (Gradient-weighted Class Activation Mapping) and SHAP (SHapley Additive exPlanations) are two popular techniques used to interpret machine learning models, such as those predicting 'Cat' images. Grad-CAM provides visual explanations by highlighting important regions in the image, whereas SHAP assigns a value to each feature, indicating its contribution to the prediction. A strength of Grad-CAM is its ability to provide intuitive visualizations, while SHAP's strength lies in its ability to assign precise values to each feature. However, Grad-CAM can be limited by its reliance on gradient information, and SHAP can be computationally expensive. Combining both techniques can provide a more comprehensive understanding of the model's decision-making process, with Grad-CAM identifying relevant regions and SHAP quantifying the contribution of each feature. By interpreting both Grad-CAM and SHAP results together, a more accurate and nuanced understanding of the 'Cat' prediction can be achieved.

5. Fidelity Metrics

```
Fidelity metrics not computed yet.
```

6. Robustness Metrics

```
Robustness metrics not computed yet.
```

7. Combined Interpretation Summary

The dog_cat model's decision-making process can be understood by combining insights from Grad-CAM and SHAP. Grad-CAM highlights broader regions of the input image that contribute to the model's prediction, such as the cat's whiskers, ears, and eyes, which are emphasized due to their high activations in the convolutional layers. Meanwhile, SHAP provides a more fine-grained attribution by assigning positive or negative values to each pixel, indicating its contribution to the "Cat" classification. By integrating these perspectives, it becomes clear that the model is identifying feline-specific features, such as whiskers and pointy ears, as key drivers of its prediction, while also considering the broader context of the image. The synergy between Grad-CAM and SHAP enables a more comprehensive understanding of the model's behavior, revealing both the specific pixels and regions that influence its decision. Overall, this unified explanation suggests that the model is leveraging a combination of local and global image features to form its prediction, although the low confidence score raises concerns about the model's reliability.

Explainability Report • Auto-generated using LLM + XAI Pipeline