Model Explainability Report

Generated at: 2025-11-22T15:15:05.403928

1. Model Summary

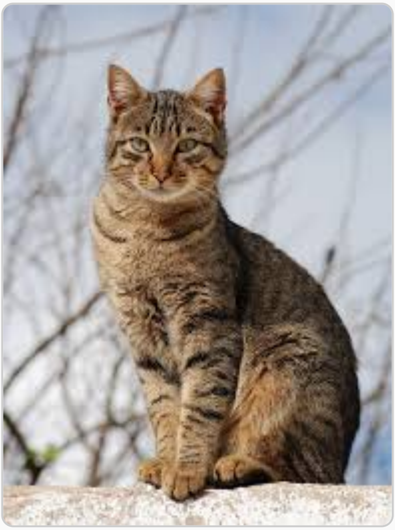
Model Key: dog_cat

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

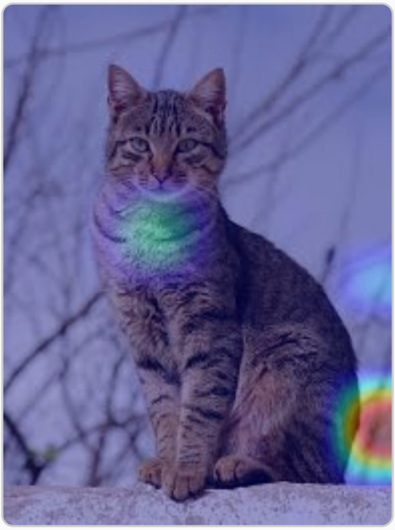
Prediction: Cat

Confidence: 0.0013

Input Image

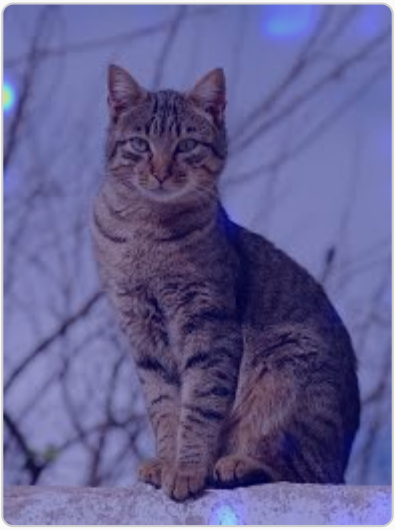


2. Grad-CAM Visualization



> 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 the cat. Furthermore, the heatmap may be highlighting noise or artifacts in the image that are not relevant to the model's prediction, 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, pixel-level attribution, whereas Grad-CAM offers a coarser, feature-level attribution. By combining these techniques, we can gain a more comprehensive understanding of the model's decision-making process, with SHAP highlighting specific pixels that drive the prediction and Grad-CAM identifying broader regions of interest. This multimodal explanation approach facilitates a more nuanced interpretation of the dog_cat 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, while SHAP assigns a value to each feature, indicating its contribution to the prediction. A strength of Grad-CAM is its ability to visualize feature importance, but it can be sensitive to hyperparameters, whereas SHAP provides more precise feature attribution but 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 explanations together, one can gain a deeper

insight into why a model predicts 'Cat' for a particular image. Overall, combining Grad-CAM and SHAP can lead to more accurate and reliable model interpretations.

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 through a combination of Grad-CAM and SHAP insights, which provide both feature-level and pixel-level attributions. The Grad-CAM heatmap 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 distinctive feline features. Meanwhile, SHAP visualization offers a more fine-grained analysis, attributing the prediction to specific pixels that either increase or decrease the likelihood of the "Cat" classification. By integrating these techniques, it becomes apparent that the model's decision is driven by a combination of feline-specific features, such as whiskers and pointy ears, and the absence of canine characteristics. The model's low confidence score, however, suggests that its decision may be unreliable, and the attributions provided by Grad-CAM and SHAP should be interpreted with caution. Overall, the unified explanation reveals that the model's prediction is based on a complex interplay of feature-level and pixel-level factors, which can be nuanced and context-dependent.

Explainability Report • Auto-generated using LLM + XAI Pipeline